SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Lucija Drašinac

IZRAČUN BIHEVIJORALISTIČKOG KREDITNOG SKORINGA ZA KLIJENTE BANKE

Diplomski rad

Voditelj rada:
Prof. dr. sc. Siniša Slijepčević

Zagreb, studeni, 2018.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Hvala mojoj dragoj prijateljici Teni koja je vjerovala u mene tijekom svih ovih godina studija i koja mi je bila najveća motivacija u teškim trenutcima. Hvala mojim roditeljima, sestri i bratu jer su bili neizmjerna podrška i jer su kroz teškoće mojih ispita prolazili sa mnom. Hvala svim prijateljima s fakulteta koji su nesebično podijelili materijale, sudjelovali na zajedničkim projektima i nerijetko bili stup podrške i oslonca. Ovdje posebnu zahvalu dugujem Magdaleni koja je sate provela učeći sa mnom, hvala mojim dragim prijateljima Luki, Rajanu, Mariju i Damjanu i hvala mojim prijateljicama Ani i Mirjam. Hvala Ivanu jer je posljednje korake studija učinio lakšima. Hvala i dragom mentoru, prof. dr. sc. Siniši Slijepčeviću na svim uputama, strpljenju i pomoći koju mi je uputio tijekom studija, a posebno tijekom pisanja diplomskog rada. Na poslijetku, hvala Bogu koji je bio nepresušni izvor mira i hrabrosti u svakom trenutku i segmentu mog hoda prema diplomi.

Sadržaj

Uvod

Kreditni skoring numerički je sustav pomoću kojeg se ocjenjuje rizičnost klijenta kojemu se želi prodati neki proizvod. Upotrebom statističkih metoda izračunava se indeks koji predstavlja vjerojatnost da će klijent biti uspješan u podmirivanju svojih obaveza. Na temelju dobivene ocjene rizičnosti, donosi se odluka o tome hoće li se proizvod prodati klijentu ili neće.

Ovaj rad sastoji se od četiri poglavlja. U prvom poglavlju navodimo osnovne definicije i rezultate koji će biti potrebni za proučavanje kreditnog skoringa, točnije bit će potrebni prilikom razvoja statističkih metoda koje izračunavaju indekse rizičnosti. Poglavlje je podijeljeno u četiri potpoglavlja. U prvom potpoglavlju definiramo slučajne varijable i pripadne funkcije distribucije. To nas dovodi do drugog potpoglavlja u kojem definiramo matematičko očekivanje i kao rezultate navodimo granične teoreme. Nakon matematičkog očekivan ja slijedi definicija i svojstva varijance i momenata, a poglavlje završavamo centralnim graničnim teoremima. Definicije i rezultati ovog poglavlja preuzeti su iz knjige [1].

Drugo poglavlje bavi se definicijom i svojstvima generaliziranog linearnog modela. On nam je potreban za definiciju linearne regresije i, u konačnici, logističke regresije kao najčešće korištene statističke metode prilikom istraživanja i razvoja kreditnog skoringa. U ovom poglavlju najprije definiramo generalizirani linearni model, a zatim radi jednostavnosti prilikom tumačenja rezultata, objašnjavamo kanonsku formu zapisivanja rezultata testiranja. Ovim poglavljem dominira Gauss-Markovljev teorem koristan za izračun najmanjih kvadratnih procjenitelja u modelu. Na kraju poglavlja objašnjavamo jednostavnu linearnu regresiju te distribucije korištene u generaliziranom linearnom modelu. Izvor ovog poglavlja može se pronaći u knjizi [2]

Glavni dio rada čini treće poglavlje jer je ono direktno vezano uz izračun kreditnog skoringa. U tom poglavlju dajemo uvod u povijesni razvoj kreditnog skoringa, a zatim opisujemo statističke metode počevši s diskriminantnom analizom. Diskriminantnom se analizom bavimo u iduća tri potpoglavlja i nakon toga u petom potpoglavlju definiramo logističku regresiju kao korisnu statističku metodu u kreditnom skoringu. U posljednjem potpoglavlju objašnjavamo neparametarsku tehniku klasifikacije klijenata u homogene skupine poznatiju kao stabla odlučivanja. Tu navodimo indekse kao vjerojatnosti da je klijent

pouzdan u podmirivanju svojih obaveza. Teoriju ovog poglavlja pronalazimo u knjizi [3] i znanstvenom radu [4].

U posljednjem poglavlju pozivamo se na znanstveni rad autora Domjana Barića, Marca Gaudarta, Siniše Slijepčevića i Tonija Vlaića. Oni se bave proučavanjem utjecaja direktive o izmijenjenim uslugama plaćanja (PSD2) na kreditni skoring i odluku o kreditiranju. Ukratko objašnjavamo cilj projekta, tijek projekta i rezultate. Izvor ovih podataka nalazi se u znanstvenom radu [5].

Poglavlje 1

Opća teorija vjerojatnosti

1.1 Slučajne varijable

Definicija i osnovna svojstva slučajnih varijabli

Osnovni pojam u teoriji vjerojatnosti jest **vjerojatnosni prostor** $(\Omega, \mathcal{F}, P)$. On nam služi kao matematički model za proučavanje slučajnih pokusa. U vezi sa slučajnim pokusima najčešće provodimo mjerenja, tj. svakom rezultatu slučajnog pokusa pridružujemo neki realan broj. Dakle, važno je promatrati realne funkcije na Ω koje ćemo zvati **slučajne varijable**. Stoga ćemo u ovom poglavlju, osim vjerojatnosnog prostora, u matematičkom smislu definirati i **slučajne varijable**. Da bismo u slučaju općeg vjerojatnosnog prostora $(\Omega, \mathcal{F}, P)$ razvili matematičku teoriju vjerojatnosti, potrebno je u skup svih slučajnih varijabli definiranih na Ω uvesti matematičku strukturu. Konkretno, nas će zanimati **neprekidne slučajne varijable** pa ćemo u daljnjem tekstu objasniti svojstva neprekidnih slučajnih varijabli i navesti neke najpoznatije primjere.

Prije svega, za definiciju neprekidnih slučajnih varijabli potrebni su nam pojmovi s Teorije mjere te se stoga u daljnjem tekstu prisjećamo potrebnih definicija.

Definicija 1.1.1. *Familija $\mathcal{F}$ podskupova od Ω jest σ-**algebra skupova** ako je*

F1. $\emptyset \in \mathcal{F}$
F2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
F3. $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.1.2. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređeni par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.*

Definicija 1.1.3. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $P : \mathcal{F} \to \mathbb{R}$ jest* ***vjerojatnost*** *ako vrijedi*

P1. $P(A) \geq 0, A \in \mathcal{F}; P(\Omega) = 1$
P2. $A_i \in \mathcal{F}, i \in \mathbb{N}\ i\ A_i \cap A_j = \emptyset\ za\ i \neq j$

Definicija 1.1.4. *Uređena trojka $(\Omega, \mathcal{F}, P)$, gdje je $\mathcal{F}$ σ-algebra na Ω i P vjerojatnost na $\mathcal{F}$, zove se* ***vjerojatnosni prostor****.*

Definicija 1.1.5. *Neka je $\mathbb{R}$ skup realnih brojeva. S $\mathcal{B}$ označimo σ-algebru generiranu familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo σ-**algebra Borelovih skupova na** $\mathbb{R}$, a elemente σ-algebre $\mathcal{B}$ zovemo* ***Borelovi skupovi****.*

Iz definicije slijedi da je svaki otvoreni interval $(a, b), a, b \in \mathbb{R}$ Borelov skup. Također, svaki zatvoreni interval $[a, b]$ je Borelov skup kao komplement otvorenog skupa. Također, intervali $(a, b]$ i $[a, b)$ jesu Borelovi skupovi. Budući da su neograničeni intervali prebrojive unije ograničenih intervala, oni su također Borelovi skupovi.

Jednočlani skupovi $\{b\}, b \in \mathbb{R}$ su zatvoreni, pa su u $\mathcal{B}$. Odavde slijedi da su prebrojivi podskupovi od $\mathbb{R}$ Borelovi.

Sada kad smo se upoznali s Borelovim skupovima, možemo definirati slučajne varijable.

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ jest* ***slučajna varijabla*** *(na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$ tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.1.7. *Funkcija $g : \mathbb{R} \to \mathbb{R}$ jest* ***Borelova funkcija*** *ako je $g^{-1}(B) \in \mathcal{B}$ za svako $B \in \mathcal{B}$ tj. ako je $g^{-1}(B) \subset \mathcal{B}$.*

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i X slučajna varijabla na Ω. Za $B \in \mathcal{B}$ stavimo:

$$P_X(B) = P(X^{-1}(B)) = P\{\omega \in \Omega; X(\omega) \in B\} = P\{X \in B\}$$

Gornjom relacijom definirana je funkcija $P_X : \mathcal{B} \to [0, 1]$ i lako se provjeri da je P_X vjerojatnost, odnosno vjerojatnosna mjera na $\mathcal{B}$. P_X zovemo **vjerojatnosna mjera inducirana s** X, a vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$ zovemo **vjerojatnosni prostor induciran s** X. Prema tome, svakoj slučajnoj varijabli X preko gornje relacije na prirodan se način pridružuje vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$.

P_X često zovemo i **zakon razdiobe** od X.

Svojstva funkcije distribucije

Jedan od osnovnih pojmova u teoriji vjerojatnosti jest pojam funkcije distribucije slučajne varijable. U teoriji vjerojatnosti operacije se izvode na funkcijama distribucije slučajnih varijabli. Osnovna klasifikacija slučajnih varijabli provodi se na osnovi oblika njihovih funkcija distribucije. Stoga u ovom poglavlju definiramo funkciju distribucije i navodimo njezina svojstva.

Definicija 1.1.8. *Neka je X slučajna varijabla na Ω.* ***Funkcija distribucije od*** *X jest funkcija $F_X : \mathbb{R} \to [0,1]$ definirana s:*

$$F_X(x) = P_X((-\infty, x]) = P(X^{-1}(-\infty, x]) = P\{\omega \in \Omega; X(\omega) \leq x\} = P\{X \leq x\}, x \in \mathbb{R}$$

Koristit ćemo $F_X = F$ ukoliko bude poznato o kojoj se slučajnoj varijabli tj. njezinoj funkciji distribucije radi.

Sljedeći teorem daje osnovna svojstva funkcije distribucije slučajne varjable.

Teorem 1.1.9. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na $\mathbb{R}$ te zadovoljava*

$$F(-\infty) = \lim_{n \to +\infty} F(x) = 0$$
$$F(+\infty) = \lim_{n \to +\infty} F(x) = 1$$

Dokaz teorema može se pronaći u [1, Teorem 9.1].

Korolar 1.1.10. *Funkcija distribucije F neprekidna je u točki $x \in \mathbb{R}$ ako i samo ako je*

$$P\{\omega \in \Omega; X(\omega) = x\} = P\{X = x\} = 0.$$

Dokaz korolara može se pronaći u [1, Korolar 9.1].
Funkciju $F : \mathbb{R} \to [0,1]$ koja ima svojstva iz teorema 1.1.9. zvat ćemo **vjerojatnosna funkcija distribucije** (na R) ili kraće **funkcija distribucije.**

Navodimo teorem koji dokazuje da svaka vjerojatnosna funkcija distribucije na $\mathbb{R}$ određuje jedinstvenu vjerojatnosnu mjeru na $\mathcal{B}$.

Teorem 1.1.11. *Neka je $F : \mathbb{R} \to [0,1]$ vjerojatnosna funkcija distribucije. Tada postoji vjerojatnosna mjera $P = P_F$ na $\mathcal{B}$ koja je jednoznačno određena s F pomoću*

$$P_F((-\infty, x]) = F(x), x \in \mathbb{R}.$$

Dokaz teorema može se pronaći u [1, Teorem 9.2].

Neprekidne slučajne varijable

U teoriji vjerojatnosti uglavnom se promatraju, tj. u primjenama se pojavljuju diskretne i neprekidne slučajne varijable. Može se pokazati da postoje slučajne varijable koje nisu niti diskretne, niti neprekidne, no u ovom radu njima se nećemo baviti. Potrebe našeg rada ograničit će se na neprekidnim slučajnim varijablama pa ćemo u podpoglavlju definirati pojam neprekidne slučajne varijable i dati najvažnije primjere.

Definicija 1.1.12. *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ i neka je F_X njezina funkcija distribucije. Kažemo da je X* ***apsolutno neprekidna*** *ili kraće* ***neprekidna slučajna varijabla*** *ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ $(f : \mathbb{R} \to \mathbb{R}_+)$ takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t)d\lambda(t), \ x \in \mathbb{R}.$$

Za funkciju distribucije F_X neprekidne slučajne varijable X kažemo da je **apsolutno neprekidna funkcija distribucije**. Ako je X neprekidna slučajna varijabla, tada se funkcija f zove **funkcija gustoće vjerojatnosti od** X tj od njezine funkcije distribucije F_X ili kraće **gustoća od** X te ju ponekad označavamo s f_X

Propozicija 1.1.13. *Neka je $f : \mathbb{R} \to \mathbb{R}$ Borelova funkcija. Da bi f bila gustoća vjerojatnosti neke neprekidne slučajne varijable X, nužno je i dovoljno da vrijedi*

(i) $f(x) \geq 0, x \in \mathbb{R}$

(ii) $\int_{-\infty}^{+\infty} f(x)d\lambda(x) = 1.$

Dokaz propozicije može se pronaći u [1, Propozicija 9.5].

Primjeri neprekidnih slučajnih varijabli

Sada ćemo navesti primjere neprekidnih slučajnih varijabli koje se najčešće pojavljuju u primjenama.

Primjer 1.1.14. *Neprekidna slučajna varijabla X ima* ***uniformnu distribuciju*** *na segmentu $[a, b], a, b \in \mathbb{R}, a < b$ ako joj je gustoća f dana s*

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0 & x \notin [a, b]. \end{cases}$$

Primjer 1.1.15. *Neprekidna slučajna varijabla X ima* ***eksponencijalnu distribuciju*** *ako joj je gustoća f dana s*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0 & x \leq 0 \end{cases}$$

gdje je $\lambda > 0$ fiksan. λ zovemo ***parametar*** *eksponencijalne distribucije.*

Primjer 1.1.16. *Neprekidna slučajna varijabla X ima* ***dvostranu eksponencijalnu distribuciju*** *na segmentu ako joj je gustoća f dana s*

$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|},\ x \in \mathbb{R}$$

gdje je $\lambda > 0$ fiksan. λ zovemo ***parametar*** *dvostrane eksponencijalne distribucije.*

Primjer 1.1.17. *Neka su $a, b \in \mathbb{R}$ i $a > 0$. Neprekidna slučajna varijabla X ima* ***Cauchyjevu distribuciju a parametrima*** *a i b ako joj je gustoća f dana s*

$$f(x) = \frac{a}{\pi[a^2 + (x-b)^2]},\ x \in \mathbb{R}$$

X ima ***jediničnu Cauchyjevu distribuciju*** *ako je $a = 1$ i $b = 0$ tj.*

$$f(x) = \frac{1}{\pi[1 + x^2]},\ x \in \mathbb{R}$$

Primjer 1.1.18. *Neka su $m, \sigma \in \mathbb{R}, \sigma > 0$. Neprekidna slučajna varijabla X ima* ***normalnu distribuciju s parametrima*** *m i σ^2 ako joj je gustoća f dana s*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},\ x \in \mathbb{R}$$

To ćemo označavati s $X \sim N(m, \sigma^2)$. X je ***jedinična normalna distribucija*** *ako je $X \sim N(0, 1)$ tj.*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},\ x \in \mathbb{R}$$

Primjer 1.1.19. *Neka su $\alpha > 0, \beta > 0$ i $\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt, x > 0$ tj. Γ je* ***gama-funkcija****. Neprekidna slučajna varijabla X ima* ***gama-distribuciju s parametrima*** *α i β ako joj je gustoća f dana s*

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0 \\ 0 & x \leq 0. \end{cases}$$

Ako je $\alpha = \frac{n}{2}, n \in \mathbb{N}$ *i* $\beta = 2$*, tada kažemo da X ima* χ^2***-distribuciju*** *s parametrom n, što često označavamo* $X \sim \chi^2(n)$ *pri čemu n zovemo* ***broj stupnjeva slobode*** *od X. Funkcija gustoće* χ^2*-distribucije s n stupnjeva slobode jest*

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}, & x > 0 \\ 0 & x \leq 0. \end{cases}$$

Primjer 1.1.20. *Neka je* $n \in \mathbb{N}$*. Neprekidna slučajna varijabla X ima* ***Studentovu t-distribuciju*** *s n* ***stupnjeva slobode*** *(oznaka je* $X \sim t(n)$*) ako joj je gustoća f dana s*

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{1}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, x \in \mathbb{R}.$$

Primjer 1.1.21. *Za* $x, y > 0$ *neka je*

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt,$$

tj. B je ***beta-funkcija****. Neka su* $p > 0, q > 0$ *fiksni. Neprekidna slučajna varijabla X ima* ***beta-distribuciju s parametrima*** *p i q ako joj je gustoća f dana s*

$$f(x) = \begin{cases} \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}, & 0 < x < 1 \\ 0 & x \leq 0 \text{ ili } x \geq 1. \end{cases}$$

1.2 Matematičko očekivanje

Definicija i osnovna svojstva matematičkog očekivanja

U ovom ćemo poglavlju uvesti pojam matematičkog očekivanja slučajnih varijabli definiranih na općem vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Navest ćemo nekoliko graničnih teorema koji se često koriste u teoriji vjerojatnosti.
Definicija matematičkog očekivanja provodi se u tri koraka. Prvo se definira matematičko očekivanje jednostavne slučajne varijable, zatim nenegativne slučajne varijable i na kraju opće slučajne varijable.
Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. S $\mathcal{K}$ označimo skup svih jednostavnih slučajnih varijabli definiranih na Ω, a s $\mathcal{K}_+$ skup svih nenegativnih funkcija iz $\mathcal{K}$.
Neka je $X \in \mathcal{K}, X = \sum_{k=1}^{n} x_k K_{A_k}$ gdje su $A_1, ..., A_n \in \mathcal{F}$ međusobno disjunktni.

Definicija 1.2.1. ***Matematičko očekivanje od*** *X ili kraće* ***očekivanje od*** *X koje označavamo s* $\mathbb{E}X$ *definira se s*

$$\mathbb{E}X = \sum_{k=1}^{n} x_k P(A_k).$$

Propozicija 1.2.2. *(i) Neka je* $c \in \mathbb{R}$ *i* $X \in \mathcal{K}$*. Tada je* $\mathbb{E}(cX) = c\mathbb{E}X$*.*

(ii) Za $X, Y \in \mathcal{K}$ *vrijedi* $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$*.*

(iii) Neka su $X, Y \in \mathcal{K}$ *i* $X \leq Y$*. Tada je* $\mathbb{E}X \leq \mathbb{E}Y$*.*

Dokaz propozicije može se pronaći u [1, Propozicija 10.1].

Definicija 1.2.3. ***Matematičko očekivanje od*** *X ili kraće* ***očekivanje od*** *X definira se s*

$$\mathbb{E}X = \lim_{n \to +\infty} \mathbb{E}X_n$$

Za nenegativnu slučajnu varijablu X vrijedi

$$\mathbb{E}X = \sup\{\mathbb{E}Y; Y \in \mathcal{K}_+, Y \leq X\}$$

.

Očito vrijedi: Ako je $X \geq 0$ tada je $\mathbb{E}X \geq 0$. To svojstvo zovemo **pozitivnost matematičkog očekivanja.**
Neka je sada X proizvoljna slučajna varijabla na Ω. Vrijedi $X = X^+ - X^-$, gdje su X^+, X^- slučajne varijable i $X^+, X^- \geq 0$.

Definicija 1.2.4. *Kažemo da* ***matematičko očekivanje od*** *X koje označavamo s* $\mathbb{E}X$ ***postoji*** *ili da je* ***definirano*** *ako je barem jedna od veličina* $\mathbb{E}X^+$ *ili* $\mathbb{E}X^-$ *konačna tj. vrijedi*

$$\min\{\mathbb{E}X^+, \mathbb{E}X^-\} < +\infty$$

.

Tada je po definiciji

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-.$$

U teoriji vjerojatnosti za matematičko očekivanje često se koristimo oznakama

$$\mathbb{E}X = \int_\Omega X dP = \int_\Omega X(\omega) dP(\omega) = \int_\Omega X(\omega) P(d\omega).$$

Matematičko očekivanje slučajne varijable X **konačno** je ako je $\mathbb{E}X^+ < +\infty$ i $\mathbb{E}X^- < +\infty$
Navodimo teorem koji daje osnovna svojstva matematičkog očekivanja.

Teorem 1.2.5. *(i) Ako $\mathbb{E}X$ postoji i $c \in \mathbb{R}$, tada $\mathbb{E}(cX)$ postoji i vrijedi*

$$\mathbb{E}(cX) = c\mathbb{E}(X).$$

(ii) Ako $X \leq Y$, tada je

$$\mathbb{E}X \leq \mathbb{E}Y$$

u smislu da

$$\text{ako je } -\infty < \mathbb{E}X, \text{ tada je } -\infty < \mathbb{E}Y \text{ i } \mathbb{E}X \leq \mathbb{E}Y$$

ili

$$\text{ako je} \mathbb{E}Y < \infty, \text{ tada je} \mathbb{E}X < \infty \text{ i } \mathbb{E}X \leq \mathbb{E}Y.$$

(iii) Ako $\mathbb{E}X$ postoji, tada je

$$|\mathbb{E}X| \leq \mathbb{E}|X|.$$

(iv) Ako $\mathbb{E}X$ postoji, tada postoji $\mathbb{E}(XK_A)$za svako $A \in \mathcal{F}$. Ako je $\mathbb{E}X$ konačno, tada je $\mathbb{E}(XK_A)$ konačno za svako $A \in \mathcal{F}$.

(v) Neka su X i Y nenegativne slučajne varijable ili $X, Y \in \mathcal{L}(P)$. Tada vrijedi

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y.$$

Dokaz teorema može se pronaći u [1, Teorem 10.1].

Granični teoremi za matematičko očekivanje

Vrlo važni teoremi u teoriji vjerojatnosti i statistici jesu granični teoremi za matematičko očekivanje. Vrlo važnu primjenu imaju i u teoriji mjere, a u ovom poglavlju navest ćemo najvažnije.

Teorem 1.2.6. *(Lebesgueov teorem o monotonoj konvergenciji)*
Neka je $(X_n, n \in \mathbb{N})$ rastući niz nenegativnih slučajnih varijabli i neka je (g.s.) $\lim_{n\to+\infty} X_n = X$. Tada je

$$\lim_{n\to+\infty} \mathbb{E}X_n = \mathbb{E}X.$$

Dokaz teorema može se pronaći u [1, Teorem 10.2].

Korolar 1.2.7. *Neka je $(X_n, n \in \mathbb{N})$ niz nenegativnih slučajnih varijabli. Tada je*

$$\mathbb{E}\Big(\sum_{n=1}^{+\infty} X_n\Big) = \sum_{n=1}^{+\infty} \mathbb{E}X_n$$

Dokaz korolara može se pronaći u [1, Korolar 10.1].

Teorem 1.2.8. *(Fatou)*
Neka je $(X_n, n \in \mathbb{N})$ *niz nenegativnih slučajnih varijabli i neka je (g.s.)* $\liminf_{n\to+\infty} X_n = X$. *Tada je*

$$\mathbb{E}X \leq \liminf_{n\to+\infty} \mathbb{E}X_n.$$

Dokaz teorema može se pronaći u [1, Teorem 10.3].

Teorem 1.2.9. *(Lebesgueov teorem o dominiranoj konvergenciji)*
Neka je $(X_n, n \in \mathbb{N})$ *niz slučajnih varijabli takav da je (g.s.)* $\lim_{n\to+\infty} X_n = X$ *i neka je* $|X_n| \leq Y$ *(g.s.) za sve n, pri čemu je* $Y \in \mathcal{L}(P)$. *Tada je*

$$\lim_{n\to+\infty} \mathbb{E}X_n = \mathbb{E}X.$$

Dokaz teorema može se pronaći u [1, Teorem 10.6].

1.3 Varijanca i momenti

Definicija i osnovna svojstva varijance i momenta

U vjerojatnosti i statistici, matematičko očekivanje i varijanca najvažnije su numeričke značajke slučajnih varijabli. U ovom ćemo poglavlju definirati **varijancu** i **momente** te ćemo navesti neke važne nejednakosti čija je primjena u statistici jako velika.

Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ i $r > 0$.

Definicija 1.3.1. $\mathbb{E}(X^r)$ *zovemo r-**ti moment od** X, a* $\mathbb{E}(|X|^r)$ *zovemo r-**ti apsolutni moment od** X.*

Definicija 1.3.2. *Neka* $\mathbb{E}X$ *postoji (tj. konačno je). Tada* $\mathbb{E}[(X-\mathbb{E}X)^r]$ *zovemo r-**ti apsolutni centralni moment od** X, a* $\mathbb{E}[|X - \mathbb{E}X|^r]$ *zovemo r-**ti apsolutni centralni moment od** X.*

Definicija 1.3.3. ***Varijanca od** X koju označujemo s* $\operatorname{Var} X$ *ili* σ_X^2 *jest drugi centralni moment od X tj.*

$$\operatorname{Var} X = \mathbb{E}[(X - \mathbb{E}X)^2]$$

*Pozitivan drugi korijen iz varijance zovemo **standardna devijacija od** X i označujemo* σ_X.

Iz definicije slijedi da je varijanca mjera odstupanja slučajne varijable X od njezina matematičkog očekivanja te da je $\operatorname{Var} X \geq 0$.

Propozicija 1.3.4. *(i) Ako je $r > 0$ i $\mathbb{E}(X^r)$ konačno, tada je $\mathbb{E}(X^s)$ konačno za $0 \leq s < r$.*

(ii) Neka X ima varijancu i neka su $a, b \in \mathbb{R}$. Tada vrijedi

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var} X$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.7].

Važne nejednakosti

Sljedeće nejednakosti imaju važnu ulogu u vjerojatnosti i statistici.

Propozicija 1.3.5. *Neka je X slučajna varijabla i g nenegativna Borelova funkcija takva da je $\mathbb{E}[g(X)] < +\infty$. Ako je g parna funkcija i neopadajuća na $[0, +\infty]$, tada za svako $\varepsilon > 0$ vrijedi*

$$P\{|X| \geq \varepsilon\} \leq \frac{\mathbb{E}[g(X)]}{g(\varepsilon)}.$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.8].

Korolar 1.3.6. *(Markovljeva nejednakost)*
Neka je $r > 0$ i $\mathbb{E}(|X|^r) < +\infty$. Tada za proizvoljno $\varepsilon > 0$ vrijedi

$$P\{|X| \geq \varepsilon\} \leq \frac{\mathbb{E}[|X|^r]}{\varepsilon^r}.$$

Iskaz korolara može se pronaći u [1, Korolar 10.3].

Korolar 1.3.7. *(Čebiševljeva nejednakost)*
Neka je X slučajna varijabla s konačnim očekivanjem i varijancom. Tada za proizvoljno $\varepsilon > 0$ vrijedi

$$P\{|X - \mathbb{E}X| \geq \varepsilon\} \leq \frac{\operatorname{Var} X}{\varepsilon^2}.$$

Iskaz korolara može se pronaći u [1, Korolar 10.4].

Propozicija 1.3.8. *Neka su X i g definirani kao u propoziciji 1.3.5. i neka je (g.s.) $\sup g(X) < +\infty$. Tada za proizvoljno $\varepsilon > 0$ vrijedi*

$$P\{|X| \geq \varepsilon\} \geq \frac{\mathbb{E}[g(X)] - g(\varepsilon)}{(g.s.)\sup g(X)}.$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.9].

Propozicija 1.3.9. *(Cauchy-Schwartzova nejednakost)*
Neka su X i Y slučajne varijable takve da je $\mathbb{E}(X^2) < +\infty$ *i* $\mathbb{E}(Y^2) < +\infty$. *Tada je* $\mathbb{E}(|XY|) < +\infty$ *i vrijedi*

$$[\mathbb{E}(|XY|)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2).$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.10].

Propozicija 1.3.10. *(Hölderova nejednakost)*
Neka su $p, q \in \mathbb{R}, p > 1, q > 1$ *i* $\frac{1}{p} + \frac{1}{q} = 1$.
Neka su X i Y slučajne varijable takve da je $\mathbb{E}(|X|^p) < +\infty$ *i* $\mathbb{E}(|Y|^q) < +\infty$. *Tada je* $\mathbb{E}(|XY|) < +\infty$ *i vrijedi*

$$\mathbb{E}(|XY|) \leq [\mathbb{E}(|X|^p)]^{\frac{1}{p}}[\mathbb{E}(|Y|^q)]^{\frac{1}{q}}.$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.11].

Propozicija 1.3.11. *(Nejednakost Minkowskog)*
Neka je $1 \leq p < +\infty$ *i neka su X i Y slučajne varijable takve da je* $\mathbb{E}(|X|^p) < +\infty, \mathbb{E}(|Y|^p) < +\infty$. *Tada je* $\mathbb{E}(|X + Y|^p) < +\infty$ *i vrijedi*

$$[\mathbb{E}(|X + Y|^p)]^{\frac{1}{p}} \leq [\mathbb{E}(|X|^p)]^{\frac{1}{p}} + [\mathbb{E}(|Y|^p)]^{\frac{1}{p}}.$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.12].

Propozicija 1.3.12. *Neka je* $0 < r < +\infty$ *i* $\mathbb{E}(|X|^r) < +\infty$. *Tada je*

$$\lim_{x \to +\infty} x^r P\{|X| \geq x\} = 0.$$

Dokaz propozicije može se pronaći u [1, Propozicija 10.13].

1.4 Centralni granični teoremi

Konvergencija slučajnih varijabli

U ovom poglavlju proučavat ćemo granično ponašanje niza $(S_n, n \in \mathbb{N})$ u smislu konvergencije po distribuciji. Da bismo to mogli, potrebno je najprije definirati tipove konvergencije slučajnih varijabli.

Definicija 1.4.1. *Kažemo da niz* $(X_n, n \in \mathbb{N})$ *slučajnih varijabli* ***konvergira gotovo sigurno*** *(g.s.) prema slučajnoj varijabli X ako je*

$$P\{\omega \in \Omega; X(\omega) = \lim_{n \to +\infty} X_n(\omega)\} = 1.$$

To označujemo $(g.s.) \lim_n X_n = X$ *ili* $X_n \xrightarrow{g.s.} X (n \to +\infty)$. *Takav limes je (g.s.) jedinstven.*

Definicija 1.4.2. *Kažemo da niz* $(X_n, n \in \mathbb{N})$ *slučajnih varijabli* ***konvergira po vjerojatnosti*** *prema slučajnoj varijabli X ako za svako* $\varepsilon > 0$ *vrijedi*

$$\lim_{n\to+\infty} P\{|X_n - X| \geq \varepsilon\} = 0.$$

To označujemo $(P)\lim_n X_n = X$ *ili* $X_n \xrightarrow{P} X (n \to +\infty)$. *Takav limes je također (g.s.) jedinstven.*

Definicija 1.4.3. *Neka je* $p \leq 1 < +\infty$ *i neka je* $X_n,\ X \in L_p(\Omega)$ $(n \in \mathbb{N})$. *Kažemo da niz* $(X_n, n \in \mathbb{N})$ ***konvergira u srednjem reda*** *p prema X ako vrijedi*

$$\lim_{n\to+\infty} \mathbb{E}(|X_n - X|^p) = 0.$$

To označujemo $(m^p)\lim_n X_n = X$ *ili* $X_n \xrightarrow{m^p} X (n \to +\infty)$

Definicija 1.4.4. *Kažemo da niz* $(X_n, n \in \mathbb{N})$ *slučajnih varijabli* ***konvergira po distribuciji*** *prema slučajnoj varijabli X ako je*

$$\lim_{n\to+\infty} F_{X_n}(x) = F_x(x),\ x \in C(F_x).$$

(F_X *je funkcija distribucije od X, a* $C(F_X)$ *je skup svih točaka neprekidnosti od* F_X*.)*
To označujemo s $(\mathcal{D})lim_n X_n = X$ *ili* $X_n \xrightarrow{\mathcal{D}} X\ (n \to +\infty)$.

Klasični centralni granični teoremi

Razmatrat ćemo probleme u vezi s konvergencijom po distribuciji. Ta konvergencija je najtipičnija za teoriju vjerojatnosti jer se definira pomoću funkcija distribucije.

Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih slučajnih varijabli i neka je $S_n = \sum_{k=1}^{n} X_k (n \in \mathbb{N})$.

Teorem 1.4.5. *(Levy)*
Neka je $X_n, n \in \mathbb{N}$ *niz nezavisnih jednakodistribuiranih slučajnih varijabli s očekivanjem m i varijancom* $\sigma^2, 0 < \sigma^2 < +\infty$ *i neka je* $S_n = \sum_{k=1}^{n} X_k (n \in \mathbb{N})$. *Tada vrijedi*

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1) \text{ za } n \to +\infty.$$

Dokaz teorema može se pronaći u [1, Teorem 14.1].

Korolar 1.4.6. *(de Moivre-Laplace)*
Neka je S_n $B(n,p)(n \in \mathbb{N}, 0 < p < 1)$. *Tada vrijedi*

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{D}} N(0,1) \text{ za } n \to +\infty.$$

Dokaz korolara može se pronaći u [1, Korolar 14.1].

Teorem 1.4.7. *(Ljapunov)*
Neka je $(X_n, n \in \mathbb{N})$ *niz nezavisnih slučajnih varijabli i neka je* $S_n = \sum_{k=1}^{n} X_k$, $s_n^2 = \operatorname{Var} S_n = \sum_{k=1}^{n} \operatorname{Var} X_k (n \in \mathbb{N})$. *Pretpostavimo da je* $s_1 > 0$ *i pretpostavimo da postoji* $\delta > 0$ *takav da je* $\mathbb{E}(|X_n|^{2+\delta}) < +\infty$ *za sve n i da vrijedi*

$$\lim_{n\to+\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} \mathbb{E}[|X_k - \mathbb{E}X_k|^{2+\delta}] = 0.$$

Tada

$$\frac{S_n - \mathbb{E}S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1) \text{ za } n \to +\infty.$$

Iskaz teorema može se pronaći u [1, Teorem 14.2]. Teorem se ne dokazuje jer je posljedica Lindebergovog teorema.

Teorem 1.4.8. *(Lindeberg)*
Neka je $(X_n, n \in \mathbb{N})$ *niz nezavisnih slučajnih varijabli s konačnim varijancama i neka je* $S_n = \sum_{k=1}^{n} X_k$, $m_n = \mathbb{E}X_n$, $s_n^2 = \operatorname{Var} S_n (n \in \mathbb{N})$. *Pretpostavimo da je* $s_1 > 0$. *Ako za svako* $\varepsilon > 0$ *vrijedi*

$$\lim_{n\to+\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{\{x;|x-m_k|\geq\varepsilon s_n\}} (x-m_k)^2 dF_{X_k}(x) = 0,$$

tada

$$\frac{S_n - \mathbb{E}S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1) \text{ za } n \to +\infty$$

Dokaz teorema može se pronaći u [1, Teorem 14.3].

Definicija 1.4.9. *Ako je ispunjen Lindebergov uvjet, tada za svako* $\varepsilon > 0$ *vrijedi*

$$\lim_{n\to+\infty} \max_{1\leq k\leq n} P\Big\{\frac{|X_k - m_k|}{s_n} \geq \varepsilon\Big\} = 0,$$

i u tom slučaju kažemo da su slučajne varijable $\frac{X_k - m_k}{s_n}$ ***uniformno asimptotski zanemarive*** *(uaz) ili da čine* ***infinitezimalni sistem.***

Teorem 1.4.10. *(Lindeberg-Feller)*
Neka je $(X_n, n \in \mathbb{N})$ *niz nezavisnih slučajnih varijabli s konačnim varijancama i neka je* $S_n = \sum_{k=1}^{n} X_k, m_n = \mathbb{E}X_n, s_n^2 = \operatorname{Var} S_n (n \in \mathbb{N})$. *Lindebergov uvjet*

$$\lim_{n\to+\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{\{x;|x-m_k|\geq\varepsilon s_n\}} (x - m_k)^2 dF_{X_k}(x) = 0, \text{ za sve } \varepsilon > 0,$$

vrijedi ako i samo ako

$$\frac{S_n - \mathbb{E}S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1) \text{ za } \to +\infty \text{ i } \frac{X_k - m_k}{s_n} \text{ su uaz}$$

Dokaz teorema može se pronaći u [1, Teorem 14.4].

Poglavlje 2

Generalizirani linearni model

2.1 Definicija generaliziranog linearnog modela

Definicija GLM-a

Generalizirani linearni model uključuje mnoge poznate i korisne modele koji proizlaze iz primijenjene statistike, uključujući modele za višestruku linearnu regresiju i analizu varijanci te logističku regresiju kao najčešće korištenu statističku metodu prilikom istraživanja i razvoja kreditnog skoringa..

Najprije ćemo definirati generalizirani linearni model, a zatim objasniti kanonsku formu zapisivanja rezultata testiranja. Na kraju poglavlja objasnit ćemo jednostavnu linearnu regresiju te distribucije korištene u generaliziranom linearnom modelu. Izvor ovog poglavlja može se pronaći u knjizi [2]

Definicija 2.1.1. ***Generalizirani linearni model*** *ili kraće* ***GLM*** *možemo definirati kao*

$$Y = X\beta + \varepsilon \tag{2.1}$$

gdje je promatrani podatak Y slučajni vektor u $\mathbb{R}^n$, X je $n \times p$ matrica poznatih konstanti, β je nepoznati parametar iz $\mathbb{R}^p$, a ε je vektor slučajnih pogrešaka u $\mathbb{R}^n$.

Pretpostavimo da je $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ slučajni uzorak iz $N(0, \sigma^2)$, gdje je $\sigma > 0$ nepoznati parametar. Tada je

$$\varepsilon \sim N(0, \sigma^2 I) \tag{2.2}$$

Ponekad vrijede slabiji uvjeti poput: $\mathbb{E}\varepsilon_i = 0$ za svaki i, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ za svaki i, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ za svaki $i \neq j$. U matričnoj notaciji to izgleda ovako: $\mathbb{E}\varepsilon = 0$ i $\mathrm{Cov}(\varepsilon) = \sigma^2 I$.
U slučaju kad je $Y + \varepsilon$ vektor konstanti i $\mathbb{E}\varepsilon = 0$, tada je $\mathbb{E}Y = X\beta$ i $\mathrm{Cov}(Y) = \mathrm{Cov}(\varepsilon) = \sigma^2 I$.
Ako je ε normalno distribuiran kao u (2.2), tada je

$$Y \sim N(X\beta, \sigma^2 I). \tag{2.3}$$

Primjer 2.1.2. *(Kvadratna regresija)*
U kvadratnoj regresiji, varijabla Y modelirana je kao kvadratna funkcija neke eksplanatorne[1] *varijable x i slučajne pogreške. Neka je*

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i \quad i = 1, ..., n.$$

Eksplanatorne varijable $x_1, x_2, ..., x_n$ *poznate su konstante,* β_1, β_2 *i* β_3 *nepoznati su parametri,* $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ *nezavisne su jednakodistribuirane varijable iz* $N(0, \sigma^2)$*. Ako definiramo matricu X kao*

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

tada je $Y = X\beta + \varepsilon$*.*

Primjer 2.1.3. *(Jednofaktorska analiza varijance)*
Pretpostavimo da imamo nezavisne slučajne varijable iz tri normalno distribuirane populacije sa zajedničkom varijancom σ^2 *i*

$$Y_i \sim \begin{cases} N(\beta_1, \sigma^2), & i = 1, ..., n_1; \\ N(\beta_2, \sigma^2), & i = n_1 + 1, ..., n_1 + n_2; \\ N(\beta_3, \sigma^2), & i = m_1 + n_2 + 1, ..., n_1 + n_2 + n_3 \end{cases} \overset{\text{def}}{=} n.$$

Ako definiramo

$$X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}$$

tada je $\mathbb{E}Y = X\beta$ *i model je distribuiran kao u (2.3)*

[1] opisuje međusobnu povezanost varijable s faktorom

U primjenama, parametri $\beta_1, \beta_2, ..., \beta_p$ često nastaju prirodno formiranjem modela. Kao posljedica toga, lako ih je tumačiti.
No, zbog tehničkih razloga češće se susrećemo s nepoznatim očekivanjem od Y

$$\xi \stackrel{\text{def}}{=} \mathbb{E}Y = X\beta$$

u $\mathbb{R}^n$ kao nepoznati parametar. Ako su $c_1, c_2, ..., c_p$ stupci matrice X,tada

$$\xi = X\beta = \beta_1 c_1 + ... + \beta_p c_p,$$

što implicira da ξ mora biti linearna kombinacija stupaca matrice X. Dakle, ξ mora ležati u vektorskom prostoru

$$\omega \stackrel{\text{def}}{=} [c_1, ..., c_p] = \{X\beta : \beta \in \mathbb{R}^p\}.$$

Koristeći ξ umjesto β, vektor nepoznatih parametara postaje $\theta = (\xi, \sigma)$ i poprima vrijednosti u $\Omega = \omega \times (0, +\infty)$.

Budući da Y ima očekivanje ξ, prilično je intuitivno da podaci kojima raspolažemo moraju pružiti različite informacije između bilo koje dvije vrijednosti za ξ. Vrijedi li to za β, ovisi o rangu r matrice X. Budući da X ima p stupaca, rang r je najviše stupnja p. Ako je rang od X jednak p, onda je svaka vrijednost $\xi \in \Omega$ slika jedinstvene vrijednosti $\beta \in \mathbb{R}^p$. No ako su stupci matrice X linearno zavisni, tada je netrivijalna linearna kombinacija stupaca u X jednaka 0. Stoga je $Xv = 0$ za svaki $v \neq 0$. Tada je

$$X(\beta + v) = X\beta + Xv = X\beta$$

i parametri β i $\beta^* = \beta + v$ daju isto očekivanje ξ. Y nam u ovom slučaju ne pruža nikakve informacije o razlikama parametara β i β^*.

Primjer 2.1.4. *Neka je*

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

Stupci od X zavisni jer je prvi stupac zbroj drugog i trećeg. Očito je rang matrice X jednak $r = 2$ *i to je manje od* $p = 3$. *Možemo uočiti da vrijednosti*

$$\beta = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad i \quad \beta^* = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

oba daju

$$\xi = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Kanonska forma

Rezultati testiranja i procjene u generaliziranom linearnom modelu puno se lakše tumače kad su podaci prikazani u kanonskoj formi. Neka $v_1, v_2, ..., v_n$ čine ortonormiranu bazu za $\mathbb{R}^n$ tako da razapinju ω. Vektor Y možemo zapisati kao linearnu kombinaciju vektora baze:

$$Y = Z_1 v_1 + ... + Z_n v_n \tag{2.4}$$

gdje u vektor Z spremamo koeficijente $Z_1, ...Z_n$.
Algebarski, Z se može naći uvođenjem $n \times n$ matrice O čiji su stupci vektori $v_1, ..., v_n$. Tada je O ortogonalna matrica, tj.

$$O'O = OO' = I$$

i vrijedi

$$Z = O'Y \text{ i } Y = OZ.$$

Zbog $Y = \xi + \varepsilon \Rightarrow Z = O'(\xi + \varepsilon) = O'\xi + O'\varepsilon$. Ako definiramo $\eta = O'\xi$ i $\varepsilon* = O'\varepsilon$, tada

$$Z = \eta + \varepsilon * .$$

Zbog $\mathbb{E}\varepsilon* = \mathbb{E}O'\varepsilon = O' \operatorname{Cov}(\varepsilon)O = O'(\sigma^2 I)O = \sigma^2 O'O = \sigma^2 I$, slijedi

$$\varepsilon* \sim N(0, \sigma^2 I)$$

i $\varepsilon_1^*, ..., \varepsilon_n^*$ su nezavisne jednakodistribuirane slučajne varijable iz $N(0, \sigma^2)$. Zbog $Z = \eta + \varepsilon*$,

$$Z \sim N(\eta, \sigma^2 I). \tag{2.5}$$

Nadalje, neka su $c_1, ..., c_p$ stupci matrice X. Tada $\xi = X\beta = \sum_{i=1}^{p} \beta_i c_i$ i

$$\eta = O'\xi = \begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix} \sum_{i=1}^{p} \beta_i c_i = \begin{pmatrix} \sum_{i=1}^{p} \beta_i v_1' c_i \\ \vdots \\ \sum_{i=1}^{p} \beta_i v_n' c_i \end{pmatrix}.$$

Zbog toga što $c_1, ..., c_p$ svi leže u ω, i $v_{r+1}, ..., v_n$ svi leže u $\omega^\perp$, vrijedi $v_k' c_i = 0$ za $k > r$ i

$$\eta_{r+1} = ... = \eta_n = 0. \tag{2.6}$$

Zbog $\eta = O'\xi$,

$$\xi = O\eta = (v_1 ... v_n) \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{i=1}^{r} \eta_i v_i$$

Ova formula uspostavlja relaciju između $\xi \in \omega$ i $(\eta_1, ..., \eta_r) \in \mathbb{R}^r$. Zbog $Z \sim N(\eta, \sigma^2 I)$, varijable $Z_1, ...Z_n$ su nezavisne i vrijedi $Z_i \sim N(\eta_i, \sigma^2)$. Gustoća od Z, uz činjenicu da je $\eta_{r+1} = ... = \eta_n = 0$ jest:

$$\frac{1}{\sqrt{2\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{r}(z_i - \eta_i)^2 - \frac{1}{2\sigma^2}\sum_{i=r+}^{n} z_i^2\right] = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{r}\eta_i z_i - \sum_{i=1}^{r}\frac{\eta_i^2}{2\sigma^2} - \frac{n}{2}log(2\pi\sigma^2)\right]$$

Ove gustoće čine eksponencijalnu familiju punog ranga $(r + 1)$ s potpunom dovoljnom statistikom

$$\left(Z_1, ..., Z_r, \sum_{i=1}^{n} Z_i^2\right) \tag{2.7}$$

Procjena parametara

Koristeći kanonsku formu, mnoge parametre lako je procijeniti. Zbog $\mathbb{E}Z_i = \eta_i, i = 1, ..., r, Z_i$ je nepristrani procjenitelj uniformno minimalne varijance tj. od $\eta_i, i = 1, ..., r$. Zbog $\xi = \sum_{i=1}^{r} \eta_i v_i$

$$\hat{\xi} = \sum_{i=1}^{r} Z_i v_i \tag{2.8}$$

je prirodni procjenitelj od ξ. Nadalje, zbog

$$\mathbb{E}\hat{\xi} = \sum_{i=1}^{r} \mathbb{E}Z_i v_i = \sum_{i=1}^{r} \eta_i$$

$\hat{\xi}$ je nepristran. Budući da je to funckija potpune dovoljne statistike, na neki bi način trebala biti optimalna. Jedna mjera optimalnosti mogla bi biti očekivana kvadratna udaljenost od prave vrijednosti ξ. Ako je $\hat{\xi}$ kompetentan nepristran procijenitelj, tada

$$\mathbb{E}||\tilde{\xi} - \xi||^2 = \sum_{j=1}^{n} \mathbb{E}(\tilde{\xi}_j - \xi_j)^2 = \sum_{j=1}^{n} \mathrm{Var}(\tilde{\xi}_j). \tag{2.9}$$

Zbog toga što je $\hat{\xi}_j$ napristran za ξ_j i funkcija je potpune dovoljne statistike, $\text{Var}(\hat{\xi}_j) \leq \text{Var}(\tilde{\xi}_j)$, $j = 1, ..., n$. Tako $\hat{\xi}$ minimizira svaki izraz u sumi varijanci u (2.8) i stoga

$$\mathbb{E}\|\hat{\xi} - \xi\|^2 \leq \mathbb{E}\|\tilde{\xi} - \xi\|^2.$$

Iz (2.4) slijedi da Y možemo zapisati kao

$$Y = \sum_{i=1}^{r} Z_i v_i + \sum_{i=r+1}^{n} Z_i v_i = \hat{\xi} + \sum_{i=r+1}^{n} Z_i v_i.$$

U ovom izrazu, prvi sumand, $\hat{\xi}$, leži u ω, a drugi, $Y - \hat{\xi} = \sum_{r+1}^{n} Z_i v_i$, leži u $\omega^\perp$. Ta razlika $Y - \hat{\xi}$ naziva se **vektor reziduala** i označava s e

$$e \stackrel{\text{def}}{=} Y - \hat{\xi} = \sum_{i=r+1}^{n} Z_i v_i. \tag{2.10}$$

Zbog $Y = \hat{\xi} + e$ iz Pitagorinog teorema, ako je $\tilde{\xi}$ bilo koja točka u ω, tada

$$\|Y - \tilde{\xi}\|^2 = \|\hat{\xi} - \tilde{\xi} + e\|^2 = \|\hat{\xi} - \tilde{\xi}\|^2 + \|e\|^2,$$

zato što $\hat{\xi} - \tilde{\xi} \in \omega$ je ortogonalan na $e \in \omega^\perp$. Iz formule se može zaključiti da je $\hat{\xi}$ jedinstvena točka u ω najbliža vektoru Y. Ta najbliža točka naziva se **projekcija** od Y na ω. Relacija $Y \rightsquigarrow \hat{\xi}$ je linearna i može se prikazati pomoću $n \times n$ matrice P,

$$\hat{\xi} = PY,$$

i P zovemo **(ortogonalna) projekcija matrice** na ω. Zbog $\hat{\xi} \in \omega$ vrijedi $P\hat{\xi} = \hat{\xi}$, i $P^2 Y = P(PY) = P\hat{\xi} = \hat{\xi} = PY$. Kako Y može poprimiti proizvoljne vrijednosti iz $\mathbb{R}^n$, slijedi da je $P^2 = P$. (Matrice s ovim svojstvom nazivaju se idempotentne matrice). Koristeći ortonormirane baze, P možemo zapisati kao $P = v_1 v_1' + ... + v_r v_r'$. Za eksplicitni izračun u cilju nam je koristiti formule koje ne ovise o vektorima $v_1, ..., v_r$ i o tome će biti riječ u nastavku.
Budući da se proizvoljne točke iz ω mogu zapisati kao $X\beta$ za neki $\beta \in \mathbb{R}^p$, ako je $\hat{\xi} = X\hat{\beta}$, tada β mora minimizirati

$$\|Y = X\beta\|^2 = \sum_{i=1}^{n} [Y_i - (X\beta)_i]^2 \tag{2.11}$$

za $\beta \in \mathbb{R}^p$. Iz tog razloga, $\hat{\beta}$ nazivamo **najmanji kvadratni procjenitelj** od β. Naravno, ako je rang r od X manji od p, $\hat{\beta}$ nije jedinstven. Kako bilo, sve parcijalne derivacije kriterija najmanjih kvadrata moraju iščeznuti kod $\beta = \hat{\beta}$. To često olakšava izračun $\hat{\beta}$ i $\hat{\xi}$. Drugi

pristup eksplicitnom izračunu proizlazi izravno iz geometrijskih razmatranja. Kako stupci $c_i, i = 1, ..., p,$ od X leže u ω i $e = Y - \hat{\xi}$ leži u $\omega^{\perp}$, mora postojati $c_i' e = 0$, što implicira

$$X'e = 0.$$

Kako je $Y = \hat{\xi} + e$,

$$X'Y = X'(\hat{\xi} + e) = X'\hat{\xi} + X'e = X'\hat{\xi} = X'X\hat{\beta}. \tag{2.12}$$

Ako je $X'X$ invertibilno, tada imamo

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.13}$$

Matrica $X'X$ je invertibilna ako je X punog ranga, tj $r = p$. Osim toga $X'X$ je pozitivno definitna. Kako bismo to vidjeli, neka je v svojstveni vektor od $X'X$ tako da je $\|v\| = 1$ i neka je svojstvena vrijednost jednaka λ. Tada je

$$\|Xv\|^2 = v'X'Xv = \lambda v'v = \lambda,$$

što mora biti strogo pozitivno jer $Xv = c_1v_1 + ... + c_pv_p$ ne može biti nula ako je X punog ranga. Kad je X punog ranga, tada

$$PY = \hat{\xi} = X\hat{\beta} = X(X'X)^{-1}X'Y,$$

i stoga se projekcijska matrica P na ω može zapisati kao

$$P = X(X'X)^{-1}X'. \tag{2.14}$$

Budući da je $\hat{\xi}$ nepristran, $a'\hat{\xi}$ je nepristrani procjenitelj od $a'\xi$. Taj je procjenitelj nepristrani procjenitelj uniformno minimalne varijance jer je $\hat{\xi}$ funkcija potpune dovoljne statistike. Zbog (2.12), $X'Y = X'\hat{\xi}$, i zbog (2.13), kad je X punog ranga, tada je

$$\hat{\beta} = (X'X)^{-1}X'\hat{\xi}.$$

Ova formula pokazuje da je $\hat{\beta}_i$ linearna funkcija od $\hat{\xi}$ i tada je $\hat{\beta}_i$ nepristrani procjenitelj uniformno minimalne varijance za β_i.

Gauss-Markovljev teorem

U ovom ćemo se poglavlju baviti generaliziranim linearnim modelom. Model ima $Y = X\beta + \varepsilon$ formu, ali ovog puta $\varepsilon_i, i = 1, ..., n$ ne mora biti slučajni uzorak iz $N(0, \sigma^2)$. Umjesto

toga, pretpostavljamo da $\varepsilon_i, i = 1, ..., n$, imaju očekivanje 0, tj. $\mathbb{E}\varepsilon_i = 0, i = 1, ..., n$; zajedničku varijancu, $\text{Var}(\varepsilon_i) = \sigma^2, i = 1, ..., n$; i varijable su nekorelirane, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. U matričnoj formi to možemo zapisati:

$$\mathbb{E}\varepsilon = 0 \quad i \quad \text{Cov}(\varepsilon) = \sigma^2 I.$$

Tada

$$\mathbb{E}Y = X\beta = \xi \quad i \quad \text{Cov}(Y) = \sigma^2 I.$$

Bilo koji procjenitelj oblika $a'Y = a_1 Y_1 + ... + a_n Y_n$, gdje je a vektor konstanti, naziva se **linearni procjenitelj**. Vrijedi:

$$\text{Var}(a'Y) = \text{Cov}(a'Y) = a' \text{Cov}(Y)a = a'(\sigma^2 I)a = \sigma^2 a'a = \sigma^2 ||a||^2. \tag{2.15}$$

Zbog $\mathbb{E}Y = \xi$, procjenitelj $a'\hat{\xi}$ je nepristran za $a'\xi$. Kako je $\hat{\xi} = PY, a'\hat{\xi} = a'PY = (Pa)'Y$ i zbog (2.15) imamo

$$\text{Var}(a'\hat{\xi}) = \sigma^2 ||Pa||^2. \tag{2.16}$$

Kako je P simetrična i vrijedi $P^2 = P$, imamo

$$\text{Cov}(\hat{\xi}) = \text{Cov}(PY) = P \,\text{Cov}(Y)P = P(\sigma^2 I)P = \sigma^2 P.$$

Kad je X punog ranga, možemo izračunati kovarijancu najmanjeg kvadratnog procjenitelja $\hat{\beta}$ od β kao

$$\text{Cov}(\hat{\beta}) = \text{Cov}((X'X)^{-1}X'Y) = (X'X)^{-1}X' \,\text{Cov}(Y)X(X'X)^{-1} = \sigma^2 (X'X)^{-1}. \tag{2.17}$$

Teorem 2.1.5. *(Gauss-Markovljev)*
Pretpostavimo

$$\mathbb{E}Y = X\beta \quad i \quad \text{Cov}(Y) = \sigma^2 I.$$

Tada je (najmanji kvadratni) procjenitelj $a'\hat{\xi}$ od $a'\xi$ nepristran i ima najmanju varijancu među svim nepristranim procjeniteljima

Dokaz. Pretpostavimo da je $\delta = b'Y$ također nepristrani procjenitelj. Zbog (2.15) i (2.16), varijance od δ i $a'\hat{\xi}$ definirane su kao:

$$\text{Var}(\delta) = \sigma^2 ||b||^2 \quad i \quad \text{Var}(a'\hat{\xi}) = \sigma^2 ||Pa||^2.$$

Ako ε dolazi iz normalne distribucije, budući da su oba procjenitelja nepristrana i $a'\hat{\xi}$ je nepristrani procjenitelj uniformno minimalne varijance, $\text{Var}(a'\hat{\xi}) \leq \text{Var}(\delta)$ ili

$$\sigma^2 ||Pa||^2 \leq \sigma^2 ||b||^2.$$

Ali formule za varijance procjenitelja ne ovise o normalnosti i zbog toga općenito vrijedi $\text{Var}(a'\hat{\xi}) \leq \text{Var}(\delta)$. □

Iako je $a'\hat{\xi}$ "najbolji" linearni procjenitelj, u nekim primjerima nelinearni procjenitelji mogu biti puno precizniji.

Primjer 2.1.6. *Pretpostavimo*

$$Y_i = \beta + \varepsilon_i, \quad i = 1, ..., n,$$

gdje su $\varepsilon_1, ..., \varepsilon_n$ *nezavisne jednakodistribuirane slučajne varijable s gustoćom*

$$f(x) = \frac{e^{\frac{-\sqrt{2}|x|}{\sigma}}}{\sigma\sqrt{2}}, \quad x \in \mathbb{R}.$$

Simetrično, $\mathbb{E}\varepsilon = 0, i = 1, ..., n$ *i*

$$\mathrm{Var}(\varepsilon_i) = \mathbb{E}\varepsilon_i^2 = 2\int_0^{+\infty} \frac{x^2 e^{-\sqrt{2}/\sigma}}{\sigma\sqrt{2}} dx = \frac{\sigma^2}{2}\int_0^{+\infty} u^2 e^{-u} du = \frac{\sigma^2}{2}\Gamma(3) = \sigma^2, \quad i = 1, ..., n.$$

$\mathrm{Cov}(Y) = \mathrm{Cov}(\varepsilon) = \sigma^2 I$, *i ako stavimo* $X = (1, ..., 1)'$, *tada je* $\mathbb{E}Y = X\beta$. *To pokazuje da su uvjeti Gauss-Markovljevog teorema zadovoljeni. Ako je* $a = n^{-1}X$, *tada je* $a'\xi = n^{-1}X'X\beta = \beta$. *Iz Gauss-Markovljevog teorema slijedi da je najbolji linearni procjenitelj od* β

$$\hat{\beta} = \frac{1}{n}X'\hat{\xi} = \frac{1}{n}X'X(X'X)^{-1}X'Y = \frac{1}{n}X'Y = \bar{Y}.$$

Varijanca ovog procjenitelja iznosi $\frac{\sigma^2}{n}$. *Najbolji procjenitelj može biti medijan uzorka,*

$$\tilde{Y} = med\{Y_1, ..., Y_n\} = \beta + med\{\varepsilon_1, ..., \varepsilon_n\}.$$

$\sqrt{n}(\tilde{Y} - \beta) \Rightarrow N(0, \frac{\sigma^2}{2})$. *Taj rezultat sugerira da*

$$\mathrm{Var}(\sqrt{n}(\tilde{Y} - \beta)) \to \frac{\sigma^2}{2}.$$

Varijable $n(\tilde{Y} - \beta)^2$ *jesu uniformno integrabilne. Kako je* $\mathrm{Var}(\sqrt{n}(\bar{Y} - \beta)) = \sigma^2$, *za velike n, varijanca od* $\tilde{Y}$ *je otprilike jednaka polovini varijance od* $\bar{Y}$.

Procjenitelj za σ^2

Ranije smo definirali, $Z_{r+1}, ..., Z_n$ su nezavisne jednakodistribuirane slučajne varijable iz $N(0, \sigma^2)$. Stoga je $\mathbb{E}Z_i^2 = \sigma^2, i = r + 1, ..., n$ i aritmetička sredina tih varijabli

$$S^2 = \frac{1}{n - r}\sum_{i=r+1}^{n} Z_i^2 \tag{2.18}$$

nepristrani je procjenitelj od σ^2. No S^2 je funkcija potpune dovoljne statistike $(Z_1, ..., Z_r, \sum_{i=1}^n Z_i^2)$ u (2.7) i S^2 je nepristrani procjenitelj uniformno minimalne varijance za σ^2. Procjenitelj S^2 može se izračunati iz duljine vektora reziduala e kojeg smo definirali u (2.10). Da bismo to vidjeli, zapišimo

$$\|e\|^2 = e'e = \Big(\sum_{i=r+1}^{n} Z_i v_i'\Big)\Big(\sum_{j=r+1}^{n} Z_j v_j\Big) = \sum_{i=r+1}^{n}\sum_{j=r+1}^{n} Z_i Z_j v_i' v_j.$$

Kako su $v_1, ..., v_n$ vektori ortonormirane baze, $v_i' v_j$ jednaki su nula kad je $i \neq j$ i jednaki su jedinici kad je $i = j$. Stoga prethodna formula postaje

$$\|e\|^2 = \sum_{i=r+1}^{n} Z_i^2, \tag{2.19}$$

i također

$$S^2 = \frac{\|e\|^2}{n-r} = \frac{\|Y - \hat{\xi}\|^2}{n-r}. \tag{2.20}$$

Kako je $\hat{\xi}$ u (2.8) funkcija od $Z_1, ..., Z_r$, i e u (2.10) je funkcija od $Z_{r+1}, ..., Z_n$, te su stoga S^2 i $\hat{\xi}$ nezavisni. Koristeći (2.19) i (2.20) te definiciju χ^2-distribucije dobivamo

$$\frac{(n-r)S^2}{\sigma^2} = \sum_{i=r+1}^{n} (Z_i/\sigma)^2 \sim \chi^2_{n-r}, \tag{2.21}$$

i $\frac{Z_i}{\sigma} \sim N(0,1)$.

Teorija koju smo upravo definirali može se primijeniti prilikom izračuna pouzdanih intervala za linearne procjenitelje. Ako je a vektor konstanti u $\mathbb{R}^n$, tada iz (2.16) standardna devijacija najmanjeg kvadratnog procjenitelja $a'\hat{\xi}$ od $a'\xi$ iznosi $\sigma\|Pa\|$. Ta se standardna devijacija prirodno procjenjuje kao

$$\hat{\sigma}_{a'\hat{\xi}} \overset{\text{def}}{=} S\|Pa\|.$$

Teorem 2.1.7. *U generaliziranom linearnom modelu gdje je* $Y \sim N(\xi, \sigma^2 I), \xi \in \omega$ *i* $\sigma^2 > 0$,

$$(a'\hat{\xi} - \hat{\sigma}_{a'\hat{\xi}} t_{\alpha/2,n-r}, a'\hat{\xi} + \hat{\sigma}_{a'\hat{\xi}} t_{\alpha/2,n-r})$$

je $(\alpha, 1-\alpha)$ *pouzdani interval za* $a'\xi$.

Dokaz. Zbog $a'\hat{\xi} \sim N(a'\xi, \sigma^2\|Pa\|^2)$,

$$\frac{a'\hat{\xi} - a'\xi}{\sigma\|Pa\|} \sim N(0,1).$$

Ova varijabla je nezavisna s $(n-r)S^2/\sigma^2$ jer su S^2 i $\hat{\xi}$ nezavisne. Koristeći definiciju t-distribucije, dobivamo

$$\frac{\frac{a'\hat{\xi}-a'\xi}{\sigma\|Pa\|}}{\sqrt{\frac{1}{n-r}\frac{(n-r)S^2}{\sigma^2}}} = \frac{a'\hat{\xi}-a'\xi}{S\,\|Pa\|} \sim t_{n-r}.$$

Vjerojatnost pokrivanja danog intervala iznosi

$$P(a'\hat{\xi}-S\,\|Pa\|t_{\alpha/2,n-r} < a'\xi < a'\hat{\xi}+S\,\|Pa\|t_{\alpha/2,n-r}) = P\left(-t_{\alpha/2,n-r} < \frac{a'\hat{\xi}-a'\xi}{S\,\|Pa\|} < t_{\alpha/2,n-r}\right) = 1-\alpha.$$

□

Kad je X punog ranga, β_i je linearni funkcional od ξ, procijenjen pomoću $\hat{\beta}_i$ s varijancom $\sigma[(X'X)^{-1}]_{ii}$. Stoga procijenjena standardna devijacija od β_i iznosi

$$\hat{\sigma}_{\hat{\beta}_i} = S\,\sqrt{[(X'X)^{-1}]_{ii}}$$

i

$$(\hat{\beta}_i - \hat{\sigma}_{\hat{\beta}_i}t_{\alpha/2,n-p}, \hat{\beta}_i + \hat{\sigma}_{\hat{\beta}_i}t_{\alpha/2,n-p}) \tag{2.22}$$

je $1-\alpha$ pouzdani interval za β_i

Jednostavna linearna regresija

Da bismo objasnili razvijene ideje, promatramo jednostavnu linearnu regresiju u kojoj je varijabla Y linearna funkcija sume nezavisne varijable x i standardne pogreške. Posebno

$$Y_i = \beta_1 + \beta_2(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, ..., n.$$

Nezavisne su varijable $x_1, ..., x_n$ uz aritmetičku sredinu $\bar{x}$ poznate konstante, β_1 i β_2 nepoznati su parametri, $\varepsilon_1, ..., \varepsilon_n$ jesu nezavisne jednakodistribuirane slučajne varijable iz $N(0, \sigma^2)$. To nam daje generalizirani linearni model s matricom

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}.$$

U parametriziranju očekivanja od Y (nazvanog **regresijska funkcija**) kao $\beta_1 + \beta_2(x-\bar{x})$, β_1 bit će interpretirana kao vrijednost regresije gdje je $x = \bar{x}$. Uočimo $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = 0$, što znači da su dva stupca od X ortogonalna. To će pojednostaviti mnoge rezultate kasnije. Na primjer, X će imati rang 2, osim ako su svi zapisi u drugom stupcu jednaki

nula, što je moguće samo kad je $x_1 = ... = x_n$. Budući da su vrijednosti matrice $X'X$ jednaki unutarnjem produktu stupaca matrice X, ta matrica i matrica $(X'X)^{-1}$ jesu dijagonalne:

$$X'X = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}$$

i

$$(X'X)^{-1} = \begin{pmatrix} 1/n & 0 \\ 0 & 1/\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}.$$

Kako je

$$X'Y = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i(x_i - \bar{x})^2 \end{pmatrix},$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i(x_i - \bar{x})/\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}.$$

Dakle,

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}. \tag{2.23}$$

Za procjenu σ^2, kako je

$$\hat{\xi} = \hat{\beta}_1 + \hat{\beta}_2(x_i - \bar{x}),$$
$$e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2(x_i - \bar{x}),$$

tada je

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

Ova formula može biti napisana na različite načine. Na primjer,

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2(1 - \hat{\rho}^2),$$

gdje je $\hat{\rho}$ **uzorački koeficijent korelacije** definiran s

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2(x_i - \bar{x})}{[\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2]^{1/2}}.$$

Ova jednadžba pokazuje da se $\hat{\rho}^2$ može promatrati kao udio varijacije od Y koji je objašnjen kao linearna relacija između Y i x. Koristeći (2.22) imamo

$$\left(\hat{\beta}_1 - \frac{S\,t_{\alpha/2,n-2}}{\sqrt{n}}, \hat{\beta}_1 + \frac{S\,t_{\alpha/2,n-2}}{\sqrt{n}}\right)$$

i to je $1-\alpha$ pouzdani interval za β_1 i

$$\left(\hat{\beta}_2 - \frac{S\,t_{\alpha/2,n-2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{\beta}_2 + \frac{S\,t_{\alpha/2,n-2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

$1-\alpha$ pouzdani interval za β_2.

Necentralne F i χ^2 distribucije

Teorija distribucije za testiranja u generaliziranom linearnom modelu oslanja se na necentralne F i χ^2 distribucije.

Definicija 2.1.8. *Ako su $Z_1, ..., Z_p$ nezavisne i $\delta \geq 0$ uz*

$$Z_1 \sim N(\delta, 1) \quad i \quad Z_j \sim N(0,1) \quad j = 2, ..., p,$$

tada $W = \sum_{i=1}^{p} Z_i^2$ ima necentralnu χ^2 distribuciju s necentralnim parametrom δ^2 i p stupnjeva slobode. To zapisujemo

$$W \sim \chi_p^2(\delta^2).$$

Lema 2.1.9. *Ako je $Z \sim N_p(\mu, I)$, tada $Z'Z \sim \chi_p^2(\|\mu\|^2)$.*

Dokaz. Neka je O ortogonalna matrica gdje je prvi redak jednak $\mu'/\|\mu\|$ pa je

$$O\mu = \tilde{\mu} = \begin{pmatrix} \|\mu\| \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

Tada

$$\tilde{Z} = OZ \sim N_p(\tilde{\mu}, I_p).$$

Iz definicije, $\tilde{Z}'\tilde{Z} = \sum_{i=1}^{p} \tilde{Z}^2 \sim \chi_p^2(\|\mu\|^2)$, i pretpostavka leme slijedi jer je

$$\tilde{Z}'\tilde{Z} = Z'O'OZ = Z'Z.$$

□

Iduća lema pokazuje da određeni kvadratni oblici za multivarijantne normalne vektore imaju necentralnu χ^2 distribuciju.

Lema 2.1.10. *Ako je $\Sigma p \times p$ pozitivno definitna matrica i ako je $Z \sim N_p(\mu, \Sigma)$, tada*

$$Z'\Sigma^{-1}Z \sim \chi_p^2(\mu'\Sigma^{-1}\mu).$$

Dokaz. Neka je $A = \Sigma^{-1/2}$, simetrični drugi korijen od Σ^{-1}. Tada je $AZ \sim N_p(A\mu, I_p)$ te je

$$Z'\Sigma^{-1}Z = (AZ)'(AZ) \sim \chi^2_p(||A\mu||^2).$$

Pretpostavka leme slijedi jer je $||A\mu||^2 = (A\mu)'(A\mu) = \mu'AA\mu = \mu'\Sigma^{-1}\mu$. □

Definicija 2.1.11. *Ako su V i W nezavisne varijable s* $V \sim \chi^2_k(\delta^2)$ *i* $W \sim \chi^2_m$, *tada je*

$$\frac{V/k}{W/m} \sim F_{k,m}(\delta^2),$$

necentralna F-distribucija *sa stupnjevima slobode k i m i necentralnim parametrom* δ^2. *Kad je* $\delta^2 = 0$ *ova se distribucija naziva F-distribucija,* $F_{k,m}$.

Testiranje hipoteza u generaliziranom linearnom modelu

U generaliziranom linearnom modelu, $Y \sim N(\xi, \sigma^2 I)$ s očekivanjem ξ u linearnom potprostoru ω dimenzije r. U ovom ćemo se poglavlju baviti testiranjem hipoteza $H_0 : \xi \in \omega_0$ i $H_1 : \xi \in \omega - \omega_0$ gdje je ω_0 $q-$dimenzionalni linearni potprostor od $\omega, 0 \leq q < r$. Nulte hipoteze ovog oblika nastanu kad β zadovoljava linearna ograničenja. Na primjer, možemo imati $H_0 : \beta_1 = \beta_2$ ili $H_0 : \beta_1 = 0$. (Slične ideje možemo koristiti za testiranje $\beta_1 = c$ ili sličnih afinih ograničenja)
Neka su $\hat{\xi}$ i $\hat{\xi}_0$ najmanji kvadratni procjenitelji za . Specijalno, $\hat{xi} = PY$ i $\hat{\xi}_0 = P_0Y$, gdje su P i P_0 matrice projekcije za ω i ω_0. Testiranje statističkih hipoteza temelji se na računanju $||Y - \hat{\xi}||$, udaljenosti između Y i ω, te $||Y - \hat{\xi}_0||$, udaljenosti između Y i ω_0. Zato što je $\omega_0 \subset \omega$, početna udaljenost mora biti manja, ali ako su udaljenosti usporedive, barem se kvalitativno H_0 može činiti adekvatnom. Testna statistika iznosi

$$T = \frac{n-r}{r-q}\frac{||Y - \hat{\xi}_0||^2 - ||Y - \hat{\xi}||^2}{||Y - \hat{\xi}||^2},$$

i nulta će hipoteza biti odbačena ako T prelazi odgovarajuću konstantu. Kako su $Y - \hat{\xi} \in \omega^\perp$ i $\hat{\xi} - \hat{\xi}_0 \in \omega$, vektori $Y - \hat{\xi}$ i $\hat{\xi} - \hat{\xi}_0$ su ortogonalni i prema Pitagorinom teoremu vrijedi

$$||Y - \hat{\xi}_0||^2 = ||Y - \hat{\xi}||^2 + ||\hat{\xi} - \hat{\xi}_0||^2.$$

Koristeći ovaj rezultat, formula za T sada postaje

$$T = \frac{n-r}{r-q}\frac{||\hat{\xi} - \hat{\xi}_0||^2}{||Y - \hat{\xi}||^2} = \frac{||\hat{\xi} - \hat{\xi}_0||^2}{(r-q)S^2}. \tag{2.24}$$

Ova testna statistika jednaka je generaliziranom testu omjera vjerodostojnosti. Kad je $r - q = 1$ test je uniformno najsnažniji nepristran, a kad je $r - q > 1$, test je najsnažniji među testovima koji zadovoljavaju ograničenja simetrije.
U daljnjem izračunu potrebna nam je distribucija od T definirana u idućem teoremu.

Teorem 2.1.12. *U generaliziranom linearnom modelu T je definiran s*

$$T \sim F_{r-q,n-r}(\delta^2),$$

gdje je

$$\delta^2 = \frac{\|\xi - P_0\xi\|^2}{\sigma^2} \tag{2.25}$$

Dokaz. Zapišimo

$$Y = \sum_{i=1}^{n} Z_i v_i$$

gdje je $v_1, ..., v_n$ ortonormirana baza izabrana tako da $v_1, ..., v_q$ razapinju ω_0 i $v_1, ..., v_r$ razapinju ω. Tada je, kao u (2.8)

$$\hat{\xi}_0 = \sum_{i=1}^{q} Z_i v_i \quad i \quad \hat{\xi} = \sum_{i=1}^{r} Z_i v_i$$

Također, kao u (2.5) i (2.6), $Z \sim N(\eta, \sigma^2 I)$ uz uvjet $\eta_{r+1} = ... = \eta_n = 0$. Kako su $v_i' v_j$ jednaki nula za $i \neq j$ i jednaki jedan za $i = j$, slijedi

$$\|Y - \hat{\xi}\|^2 = \Big\| \sum_{i=r+1}^{n} Z_i v_i \Big\|^2 = \Big(\sum_{i=r+1}^{n} Z_i v_i' \Big) \Big(\sum_{j=r+1}^{n} Z_j v_j \Big) = \sum_{i=r+1}^{n} \sum_{j=r+1}^{n} Z_i Z_j v_i' v_j = \sum_{i=r+1}^{n} Z_i^2$$

Slično,

$$\|Y - \hat{\xi}_0\|^2 = \sum_{i=q+1}^{n} Z_i^2$$

i

$$T = \frac{\frac{1}{r-q} \sum_{i=q+1}^{r} (Z_i/\sigma)^2}{\frac{1}{n-r} \sum_{i=r+1}^{n} (Z_i/\sigma)^2}$$

Z_i su nezavisne te su brojnik i nazivnik u definiciji od T također nezavisni. Zbog $Z_i\sigma \sim N(\eta_i/\sigma, 1)$, iz Leme 2.1.9. slijedi

$$\sum_{i=q+1}^{r} \left(\frac{Z_i}{\sigma} \right)^2 \sim \chi^2_{r-q}(\delta)^2,$$

gdje je

$$\delta^2 = \sum_{i=q+1}^{r} \frac{\eta_i^2}{\sigma^2}. \tag{2.26}$$

Kako je $\eta_i = 0$ za $i = r+1, ..., n$, $Z_i/\sigma \sim N(0,1)$, $i = r+1, ..., n$ i $\sum_{i=r+1}^{n}(Z_i/\sigma)^2 \sim \chi^2_{n-r}$. Stoga, prema Definiciji 2.1.11. za necentralnu distribuciju F, $T \sim F_{r-q,n-r}(\delta^2)$, gdje je δ dan formulom (2.26). Ostalo nam je pokazati da je

$$\sum_{i=q+1}^{r} \eta_i^2 = \|\xi - P_0\xi\|^2.$$

Kako je

$$\xi = \mathbb{E}\hat{\xi} = \sum_{i=1}^{r} \eta_i v_i$$

i

$$P_0\xi = \mathbb{E}P_0Y = \mathbb{E}\hat{\xi}_0 = \sum_{i=1}^{q} \eta_i v_i,$$

$$\xi - P_0\xi = \sum_{i=q+1}^{r} \eta_i v_i.$$

Tada, prema Pitagorinom teoremu,

$$\|\xi - P_0\xi\|^2 = \sum_{i=q+1}^{r} \eta_i^2,$$

što je i trebalo pokazati. □

Sustav pouzdanih intervala

Svako istraživanje na velikom skupu podataka nudi mogućnost određivanja pouzdanih intervala za brojne parametre. Ponekad pouzdani intervali neće sadržavati smislene vrijednosti i u želji da se to spriječi predloženi su sustavi pouzdanih intervala. U ovom ćemo poglavlju navesti nekoliko osnovnih ideja, počevši s Primjerom 2.1.3 (Jednofaktorska analiza varijance).
Model koji razmatramo ima

$$Y_{kl} = \beta_k + \varepsilon_{kl}, \quad 1 \le l \le c, \quad 1 \le k \le p.$$

Ovo možemo promatrati kao model za nezavisne slučajne uzorke iz p različitih normalno distribuiranih populacija s jednakim varijancama. Broj opažanja c jednak je u svim populacijama. Zapišemo li Y_{kl} kao vektor, dobivamo, baš kao u Primjeru, generalizirani linearni model. Najmanji kvadratni procjenitelj od β možemo minimizirati kao

$$\sum_{l=1}^{c}\sum_{k=1}^{p}(Y_{kl} - \beta_k)^2.$$

Parcijalna derivacija izraza iznad u odnosu na β_m iznosi

$$-2\sum_{l=1}^{c}(Y_{ml}-\beta_m)$$

i ona nestaje kad je $\beta_m = \hat{\beta_m}$ dan s

$$\hat{\beta_m} = \bar{Y}_m \overset{\text{def}}{=} \frac{1}{c}\sum_{l=1}^{c} Y_{ml}, \quad m = 1, ..., p,$$

To su najmanji kvadratni procjenitelji. Ovdje je $r = p$ i $n = pc$, pa je

$$S^2 = \frac{\|Y-\hat{\xi}\|^2}{pc-p} = \frac{1}{p(c-1)}\sum_{l=1}^{c}\sum_{k=1}^{p}(Y_{kl}-\hat{\beta}_k)^2.$$

Najmanji kvadratni procjenitelji jednaki su prosjecima različitih skupova podataka za Y_{kl}. Stoga su $\hat{\beta}_1, ..., \hat{\beta}_p$ nezavisni i vrijedi

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2/c), \quad k = 1, ..., p.$$

Također,

$$\frac{p(c-1)S^2}{\sigma^2} \sim \chi^2_{p(c-1)},$$

i S^2 je nezavisan s $\hat{\beta}$.

Za početak, pokušajmo odrediti intervale $I_1, ..., I_p$ od $\beta_1, ..., \beta_p$ s definiranom vjerojatnosti $1-\alpha$. Specijalno, želimo

$$P(\beta_k \in I_k, k = 1, ..., p) = 1-\alpha.$$

Intervali pouzdanosti prema (2.22) jednaki su

$$\left(\hat{\beta}_k - \frac{S}{\sqrt{c}}t_{\alpha/2,p(c-1)}, \hat{\beta}_k + \frac{S}{\sqrt{c}}t_{\alpha/2,p(c-1)}\right),$$

i intuitivno, mogli bismo pisati

$$I_k = \left(\hat{\beta}_k - \frac{S}{\sqrt{c}}q, \hat{\beta}_k + \frac{S}{\sqrt{c}}q\right), \quad k = 1, ..., p,$$

gdje je q prikladno odabran. Sada imamo

$$P(\beta_k \in I_k, k = 1, ..., p) = P\Big(|\hat{\beta}_k - \beta_k| < \frac{S}{\sqrt{c}}q, k = 1, ..., p\Big)$$

$$= P\left(\max_{1\leq k\leq p} \frac{|\hat{\beta}_k - \beta_k|}{S/\sqrt{c}} < q \right)$$
$$= P\left(\max_{1\leq k\leq p} \frac{|Z_k|}{\sqrt{W}} < q \right),$$

gdje je

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sigma/\sqrt{c}} \sim N(0,1), \quad k = 1, ..., p$$

i $W = S^2/\sigma^2$. Zato što su $Z_1, ..., Z_p$ i W nezavisni te $mW \sim \chi^2_m$ gdje je $m = p(c-1)$ i vjerojatnost ne ovisi o parametrima β i σ.

Definicija 2.1.13. *Ako su $Z_1, ..., Z_p$ i W nezavisne varijable i $Z_k \sim N(0,1), k = 1, ..., p$ i $mW \sim \chi^2_m$, tada*

$$\frac{\max_{1\leq k\leq p} |Z_k|}{\sqrt{W}}$$

*ima **studentiziranu distribuciju maksimalne apsolutne vrijednosti.** s parametrima p i m.*

Ako je q gornji α-kvantil ove distribucije, tada intervali $I_1, ..., I_p$ sadrže jednaku $1-\alpha$ vjerojatnost.

U praksi, puno je zanimljivije promatrati i uspoređivati populacije međusobno, nego pojedinačno računati očekivanja i stoga bi pouzdani intervali za razliku $\beta_j - \beta_i$ mogli biti zanimljivi. Sada računamo intervale $I_i j$ tako da vrijedi

$$P(\beta_j - \beta_i \in I_{ij}, \forall i \neq j) = 1 - \alpha$$

Prirodno možemo pretpostaviti da bi intervali koje tražimo mogli biti oblika

$$I_{ij} = \left(\hat{\beta}_j - \hat{\beta}_i - \frac{S}{\sqrt{c}} q, \hat{\beta}_j - \hat{\beta}_i + \frac{S}{\sqrt{c}} q \right)$$

gdje je q proizvoljan. Tada je

$$P(\beta_j - \beta_i \in I_{ij}, \forall i \neq j)$$
$$= P\left(|(\hat{\beta}_j - \beta_j) - (\hat{\beta}_i - \beta_i)| < \frac{S}{\sqrt{c}} q, \forall i \neq j \right)$$
$$= P\left(\frac{|\sqrt{c}(\hat{\beta}_j - \beta_j) - \sqrt{c}(\hat{\beta}_i - \beta_i)|}{S} < q, \forall i \neq j \right)$$
$$= P\left(\frac{|Z_j - Z_i|}{\sqrt{W}} < q, \forall i \neq j \right)$$

$$= P\left(\frac{\max_{1\leq q\leq p} Z_k - \min_{1\leq k\leq p} Z_k}{\sqrt{W}} < q\right).$$

Ova definicija je dobra zato što vjerojatnost ne ovisi o β ili σ.

Definicija 2.1.14. *Ako su $Z_1, ..., Z_p$ i W nezavisne varijable i $Z_k \sim N(0,1), k = 1, ..., p$ i $mW \sim \chi^2_m$, tada*

$$\frac{\max_{1\leq k\leq p} Z_k - \min_{1\leq k\leq p} Z_k}{\sqrt{W}}$$

*ima **studentiziranu distribuciju raspona** s parametrima p i m.*

Ako je q gornji α-kvantil ove distribucije, tada intervali $I_1, ..., I_p$ sadrže jednaku $1-\alpha$ vjerojatnost.
Derivacija sustava pouzdanih intervala oslanja se na strukturu ANOVA modela. Opći rezultati koriste Scheffeovu metodu. Ona se temelji na setu pouzdanih intervala za parametar $\psi \in \mathbb{R}^q$, uz uvjet $q \leq r$, što je linearna funkcija očekivanja od ξ i dana je s

$$\psi = A\xi = AX\beta$$

za neku $q \times n$ matricu A. Kada je X punog ranga, $\beta = (X'X)^{-1}X'\xi$ i $A = (X'X)^{-1}X'$ daju $\psi = \beta$. Postoje i druge linearne funkcije za β. Zbog $P\xi = \xi$, imamo $AP\xi = A\xi = \psi$ i zamjenom A s $A*$, ψ se ne mijenja. Tada je $A * P = APP = AP = A *$. Zamjenom A s $A*$, ako je potrebno, bez smanjenja općenitosti možemo pretpostaviti da je $A = AP$. Ovo je dobro definirano jer je najmanji kvadratni procjenitelj od ψ jednak

$$\hat{\psi} = A\hat{\xi} = APY = AY.$$

Konačno, možemo zaključiti da su retci matrice AX linearno nezavisni. Kako je AX punog ranga i $\psi = AX\beta$, zaključujemo da ψ može imati proizvoljne vrijednosti iz $\mathbb{R}^q$. Možemo primijetiti da će rang matrice AX biti manji od ranga matrice A, jer ako retci matrice A zadovoljavaju netrivijalno linearno ograničenje, $v'A = 0$, tada je i $v'AX = 0$, i retci matrice AX zadovoljavaju isto linearno ograničenje. Ako definiramo $B = AA'$, tada je B pozitivno definitna jer je

$$q \leq r(AX) \leq r(A) \leq q,$$

i to nam pokazuje da su A i AX obje punog ranga te $v'Bv = v'AA'v = \|A'v\|^2$, što je pozitivno, osim ako je $v = 0$ i A je punog ranga. Slijedi

$$\hat{\psi} \sim N(\psi, \sigma^2 B),$$

i prema Lemi 2.1.10.,

$$\frac{(\hat{\psi} - \psi)' B^{-1} (\hat{\psi} - \psi)}{\sigma^2} \sim \chi^2_p.$$

Zato što je $\hat{\psi}$ funkcija od $\hat{\xi}$ i $\hat{\xi}$ i S^2 su nezavisni, kvadratna forma je nezavisna s

$$\frac{(n-r)S^2}{\sigma^2} \sim \chi^2_{n-r}.$$

Tada prema Definiciji 2.1.11. slijedi

$$\frac{(\hat{\psi}-\psi)'B^{-1}(\hat{\psi}-\psi)/(q\sigma^2)}{S^2/\sigma^2} = \frac{(\hat{\psi}-\psi)'B^{-1}(\hat{\psi}-\psi)}{qS^2} \sim F_{q,n-r}.$$

Iz toga slijedi

$$P((\hat{\psi}-\psi)'B^{-1}(\hat{\psi}-\psi) \leq qS^2F_{\alpha,q,n-r}) = 1-\alpha.$$

Skup vrijednosti za ψ gdje se taj događaj ostvaruje jest višeznačna elipsa centrirana oko $\hat{\psi}$. Ova slučajna elipsa je $1-\alpha$ pouzdani set za ψ. Da bismo formirali sustav pouzdanih intervala, iz pouzdanog seta za elipsu, možemo uočiti

$$(\hat{\psi}-\psi)'B^{-1}(\hat{\psi}-\psi) = \|B_{-1/2}(\hat{\psi}-\psi)\|^2.$$

Tako, za bilo koji $h \in \mathbb{R}^q$ slijedi

$$h'Bh = h'B^{1/2}B^{1/2}h = \|B^{1/2}h\|^2.$$

Prema Schwarzovoj nejednakosti slijedi

$$\|B^{1/2}h\|^2\|B^{-1/2}(\hat{\psi}-\psi)\|^2 \geq [h'B^{1/2}B^{-1/2}(\hat{\psi}-\psi)]^2 = [h'(\hat{\psi}-\psi)]^2.$$

Dakle,

$$\begin{aligned}
P\Big\{[h'(\hat{\psi}-\psi)]^2 \leq qS^2h'BhF_{\alpha,q,n-r}, h \in \mathbb{R}^q\Big\} & \\
&\geq P(\|B^{1/2}h\|^2\|B^{-1/2}(\hat{\psi}-\psi)\|^2 \leq qS^2h'BhF_{\alpha,q,n-r}, \forall h \in \mathbb{R}^q) \\
&= P(\|B^{-1/2}(\hat{\psi}-\psi)\|^2 \leq qS^2F_{\alpha,q,n-r}) \\
&= 1-\alpha.
\end{aligned}$$

Uzimanjem $h = B^{-1}(\hat{\psi}-\psi)$, vjerojatnost može biti najviše $1-\alpha$ i tada govorimo o jednakosti. Kako je

$$Var(h'(\hat{\psi}-\psi)) = \sigma^2h'Bh,$$

prirodno je procijeniti s

$$\hat{\sigma}^2_{h'\psi} = S^2h'Bh.$$

Taj je izraz jednak izrazu

$$P\Big\{[h'(\hat{\psi}-\psi)]^2 \leq \hat{\sigma}^2_{h'\psi}qF_{\alpha,q,n-r}, \forall h \in \mathbb{R}^q\Big\} = 1-\alpha.$$

Stoga intervali

$$\left(h'\hat{\psi} - \hat{\sigma}_{h'\hat{\psi}}\sqrt{qF_{\alpha,q,n-r}}, h'\hat{\psi} + \hat{\sigma}_{h'\hat{\psi}}\sqrt{qF_{\alpha,q,n-r}}\right)$$

sadrže $h'\psi$ istovremeno za sve $h \in \mathbb{R}^q$, s vjerojatnošću $1-\alpha$.

Poglavlje 3

Statističke metode za izračun kreditnog skoringa

3.1 Uvod

Potreba za izračunom kreditnog skoringa prvi se put javila pedesetih godina prošloga stoljeća. Sve do danas statističke metode daleko su najkorištenije za izračun kreditnog skoringa. Prednost korištenja ovih metoda jest da korisniku dopuštaju iskorištavanje postojećih znanja o svojstvima procjenitelja uzorka, pouzdanim intervalima i omogućuju testiranje hipoteza u kontekstu kreditnog skoringa. U konačnici, statističke metode omogućuju izdvajanje nevažnih karakteristika, a zadržavanje važnih prilikom izračuna skoringa. Inicijalno, metode su bile bazirane na diskriminantnoj analizi koju je osmislio Ronald A. Fisher 1936. godine za rješavanje klasifikacijskih problema. To je dovelo do razvoja linearnog skoringa baziranog na Fisherovoj linearnoj diskriminantnoj funkciji. Fisherov pristup možemo vidjeti kao oblik linearne regresije i to nas dovodi do otkrića ostalih oblika regresije koji imaju manje restriktivne pretpostavke što garantira optimalnost i dovodi do razvoja pravila za linearni skoring. Daleko najuspješnija metoda jest logistička regresija koja se razvila iz linearne regresije - diskriminantne analize kao najučestalije statističke metode. Ostali pristupi koji su se razvili u proteklih 20 godina jesu klasifikacijsko stablo i rekurzivno particioniranje. Sve te metode koriste se u praksi za dobivanje skoringa, ali tu je još mnogo posla i eksperimentiranja oko korištenja statističkih metoda. U ovom ćemo poglavlju objasniti logističku regresiju i dat ćemo statističku pozadinu za razvoj spomenutih metoda. Najprije ćemo početi s diskriminantnom analizom. Opisat ćemo kako linearna diskriminantna funkcija dolazi kao klasifikator u tri različita pristupa problemu. Za definiranje ovih pojmova pozivamo se na knjigu [3] i znanstveni rad [4].

3.2 Diskriminantna analiza

Proces odobravanja kredita dovodi do dva ishoda - omogućuje podnositelju zahtjeva novi kredit ili odbija zadani zahtjev. Kreditni skoring pokušava pomoći pronaći najbolje pravilo koje bi se moglo primijeniti na zadane zahtjeve. Ako su moguća dva ishoda - prihvatiti ili odbiti zahtjev, onda se zahtjev može svrstati jedino u dvije klase - dobru i lošu. Dobar zahtjev je svaki onaj koji je prihvaćen od strane zajmodavca, dok je loš zahtjev odbijen od istog. Neka je $X = (X_1, ..., X_p)$ niz p slučajnih varijabli koji sadrži dostupne informacije o podnositelju zahtjeva za kredit, kako iz obrasca za prijavu, tako i putem kreditnog referentnog ureda. Koristimo riječi **varijabla** i **karakteristika** naizmjenično kako bismo definirali X_i: prvu kada želimo naglasiti slučajnu prirodu ove informacije između kandidata i drugu kad želimo objasniti kakva je informacija. Vrijednosti varijable za pojedinog kandidata označavamo $x = (x_1, ..., x_p)$. U terminologiji kreditnog skoringa, različite vrijednost x_i od X_i nazivaju se atributi dane karakteristike. Pretpostavimo da je A skup svih mogućih vrijednosti koje varijabla $X = (X_1, ..., X_p)$ može poprimiti. Cilj je pronaći pravilo prema kojem se skup A dijeli na dva podskupa A_G i A_B. U podskup A_G spremaju se vrijednosti koje su "dobre" i prihvatljive (eng. good), a u podskup A_B spremaju se vrijednosti koje su "loše" (eng. bad). Pretpostavimo za sada da je očekivani profit za svakog podnositelja jednak i označimo ga s L. Također, pretpostavimo da je nastali dug jednak za svakog podnositelja i označimo ga s D. Označimo s p_G vjerojatnost svih podnositelja koji su "dobri"i slično s p_B vjerojatnost svih podnositelja koji su "loši".
Pretpostavit ćemo da je A konačan skup različitih atributa x. Neka je $p(\mathbf{x}|G)$ vjerojatnost da "dobri" podnositelj zahtjeva sadrži atribut x. To je uvjetna vjerojatnost i dana je s

$$p(\mathbf{x}|G) = \frac{P(\text{podnositelj je "dobar" i sadrži atribut x})}{P(\text{podnositelj je "dobar"})} \tag{3.1}$$

Slično, definirajmo $P(\mathbf{x}|B)$ vjerojatnost da "loš" podnositelj zahtjeva sadrži atribut x.
Ako s $q(G|\mathbf{x})$ označimo vjerojatnost da je svaki podnositelj koji sadrži atribut x "dobar", tada je

$$q(G|\mathbf{x}) = \frac{P(\text{podnositelj sadrži atribut x i "dobar" je})}{P(\text{podnositelj sadrži atribut x})} \tag{3.2}$$

Ako je, uz to, $p(\mathbf{x})$ vjerojatnost da podnositelj sadrži atribut $\mathbf{x}$, tada, pomoću (3.1) i (3.2) dobijemo

$$P(\text{podnositelj je dobar i sadrži atribut } \mathbf{x}) = q(G|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|G)p_G. \tag{3.3}$$

To nas dovodi do Bayesovog teorema koji kaže

$$q(G|\mathbf{x}) = \frac{p(\mathbf{x}|G)p_G}{p(\mathbf{x})}. \tag{3.4}$$

Simetrično, dobijemo i

$$q(B|\mathbf{x}) = \frac{p(\mathbf{x}|B)p_B}{p(\mathbf{x})}. \tag{3.5}$$

Iz (3.4) i (3.5) dobijemo

$$\frac{q(G|\mathbf{x})}{q(B|\mathbf{x})} = \frac{p(\mathbf{x}|G)p_G}{p(\mathbf{x}|B)p_B}. \tag{3.6}$$

Očekivani troškovi po podnositelju zahtjeva ako uključimo samo one koji imaju atribute unutar skupa A_G, a odbacimo one s atributima iz skupa A_B iznose:

$$L\sum_{x\in A_B} p(\mathbf{x}|G)p_G + D\sum_{x\in A_G} p(\mathbf{x}|B)p_B = L\sum_{x\in A_B} q(G|\mathbf{x})p(\mathbf{x}) + D\sum_{x\in A_G} q(B|\mathbf{x})p(\mathbf{x}) \tag{3.7}$$

Pravilo odlučivanja koje smanjuje očekivane troškove dano je s

$$A_G = \{\mathbf{x}|Dp(\mathbf{x}|B)p_B \le Lp(\mathbf{x}|G)p_G\} = \left\{\mathbf{x}|\frac{D}{L} \le \frac{p(\mathbf{x}|G)p_G}{p(\mathbf{x}|B)p_B}\right\} = \left\{\mathbf{x}|\frac{D}{L} \le \frac{q(G|\mathbf{x})}{q(B|\mathbf{x})}\right\} \tag{3.8}$$

gdje posljednja jednakost slijedi iz (3.6). Pretpostavimo da je stopa prihvaćanja podnositelja zahtjeva jednaka a. Tada A_G zadovoljava

$$\sum_{x\in A_G} p(\mathbf{x}|G)p_G + \sum_{x\in A_B} p(\mathbf{x}|B)p_B = a. \tag{3.9}$$

Ako definiramo $b(\mathbf{x}) = p(\mathbf{x}|B)p_B$ za svaki $\mathbf{x} \in A$ i želimo pronaći skup A_G, tada možemo

$$\text{minimiziranjem} \sum_{x\in A_G} b(\mathbf{x}) = \sum_{x\in A_G}\left(\frac{b(\mathbf{x})}{p(\mathbf{x})}p(\mathbf{x})\right) \text{dolazimo do } \sum_{x\in A_G} p(\mathbf{x}) = a. \tag{3.10}$$

Koristeći Lagrangeove multiplikatore, možemo zaključiti da to mora biti skup atributa iz $\mathbf{x}$, uz uvjet $\frac{b(\mathbf{x})}{p(\mathbf{x})} \le c$, gdje je c izabran tako da je suma od $p(\mathbf{x})$ koja zadovoljava dani uvjet jednaka a.

Stoga je

$$A_G = \left\{\mathbf{x}|\frac{b(\mathbf{x})}{p(\mathbf{x})} \le c\right\} = \{\mathbf{x}|q(B|\mathbf{x}) \le c\} = \left\{\mathbf{x}|\frac{1-c}{c} \le \frac{p(\mathbf{x}|G)p_G}{p(\mathbf{x}|B)p_B}\right\}, \tag{3.11}$$

gdje druga nejednakost slijedi iz definicije od $p(x)$ i $b(x)$. Cijela analiza može se primijeniti uz pretpostavku da su karakteristike podnositelja zahtjeva neprekidne i da nisu diskretne slučajne varijable. Potrebno je zamijeniti $p(\mathbf{x}|G)$ i $p(\mathbf{x}|B)$ s $f(\mathbf{x}|G)$ i $f(\mathbf{x}|B)$ i sumu zamijeniti integralom. Stoga, ako skup A podijelimo na skupove A_G i A_B i u obzir uzmemo samo skup A_G dobivamo

$$L\int_{x\in A_B} f(\mathbf{x}|G)p_G d\mathbf{x} + D\int_{x\in A_G} f(\mathbf{x}|B)p_B d\mathbf{x}, \tag{3.12}$$

Pravilo odlučivanja koje smanjuje očekivane troškove dano je, analogno kao u (3.8)

$$A_G = \{\mathbf{x}|Df(\mathbf{x}|B)p_B \leq Lf(\mathbf{x}|G)p_G\} = \left\{\mathbf{x}|\frac{Dp_B}{Lp_G} \leq \frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)}.\right\} \tag{3.13}$$

Razlikujemo tri slučaja - **jednodimenzionalni normalni slučaj, višedimenzionalni normalni slučaj s jednakim kovarijancama, višedimenzionalni normalni slučaj s različitim kovarijacijskim matricama**.

Jednodimenzionalni normalni slučaj

Riječ je o najjednostavnijem mogućem slučaju kad postoji samo jedna neprekidna karakteristična varijabla X i njezine karakteristične funkcije jesu normalne, $f(x|G)$ s očekivanjem μ_G i varijancom σ^2 te $f(\mathbf{x}|B)$ s očekivanjem μ_B i varijancom σ^2. Tada je funckija gustoće jednaka

$$f(\mathbf{x}|G) = (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(\mathbf{x}-\mu_G)^2}{2\sigma^2}\right). \tag{3.14}$$

Višedimenzionalni normalni slučaj s jednakim kovarijancama

Puno realniji slučaj jest kad imamo p varijabli (karakteristika) i obe vrste varijabli - i one dobre i one loše, dolaze iz višedimenzionalne normalne distribucije. Neka je očekivanje za dobre karakteristike dano s μ_G, a za loše μ_B i neka je kovarijacijska matrica jednaka Σ. To znači da je $\mathbb{E}(X_i|G) = \mu_{G,i}, \mathbb{E}(X_i|B) = \mu_{B,i}$ i $\mathbb{E}(X_iX_j|G) = \mathbb{E}(X_iX_j|B) = \Sigma_{ij}$. Tada je odgovarajuća funkcija gustoće dana s

$$f(\mathbf{x}|G) = (2\pi\sigma^2)^{-p/2}(\det\Sigma)^{-1} \exp\left(\frac{-(\mathbf{x}-\mu_G)\Sigma^{-1}(\mathbf{x}-\mu_G)^T}{2}\right). \tag{3.15}$$

gdje je $(\mathbf{x} - \mu_G)$ vektor s jednim retkom i p stupaca.

Višedimenzionalni normalni slučaj s različitim kovarijacijskim matricama

Ovaj je slučaj podslučaj prethodnog. Kovarijacijske matrice za dobre i loše karakteristike nisu jednake. Možemo označiti kovarijacijsku matricu dobrih karakteristika sa Σ_G, a loših sa Σ_B i riješiti problem kao u prethodnom slučaju.

3.3 Diskriminantna analiza: Podjela u dvije grupe

U Fisherovom originalnom radu iz 1936. godine, koji je uključivao linearnu diskriminantnu funkciju, cilj je bio pronaći kombinaciju varijabli koje najbolje razdvajaju dostupna obilježja u dvije različite grupe. U kontekstu kreditnog skoringa, te dvije grupe kreirane su od strane zajmodavca kao grupa s dobrim i grupa s lošim karakteristikama. Neka je $Y = w_1X_1 + ... + w_pX_p$ linearna kombinacija karakteristika X. Jedna očita mjera razdvajanja jest kako različite vrijednosti daje očekivanje od Y za dvije grupe s dobrim tj. lošim obilježjima. Promatramo, dakle, razliku između $\mathbb{E}(Y|G)$ i $\mathbb{E}(Y|B)$ i određujemo w_i, gdje w_i maksimizira tu razliku i vrijedi $\sum_i w_i = 1$. Fisher predlaže da ako pretpostavimo da dvije grupe imaju zajedničku varijancu uzorka, onda je mjera odvajanja dana s

$$M = \frac{\text{razlika između očekivanja u dvjema grupama}}{(\text{varijanca svake grupe})^{1/2}}$$

Pretpostavimo da su m_G i m_B očekivanja dobre, tj loše grupe redom i neka je S zajednička varijanca za obje grupe. Ako je $Y = w_1X_1 + ... + w_pX_p$, tada je odgovarajuća udaljenost M jednaka

$$M = w^T \cdot \frac{m_G - m_B}{(w^T S\, w)^{1/2}} \tag{3.16}$$

To povlači da je $\mathbb{E}(Y|G) = w \cdot m_G^T, \mathbb{E}(Y|B) = w \cdot m_B^T$ i $\text{Var}(Y) = w \cdot S \cdot w^T$. Slijedi da je M maksimizirana kad je

$$\frac{m_G - m_B}{(w \cdot S \cdot w^T)^{1/2}} - \frac{(w \cdot (m_G - m_B)^T)(S\, w)^T}{(w \cdot S \cdot w^T)^{3/2}} = 0 \tag{3.17}$$

tj.

$$(m_G - m_B)(w \cdot S \cdot w^T) = (S\, w)^T (w \cdot (m_G - m_B)^T). \tag{3.18}$$

3.4 Diskriminantna analiza: Oblik linearne regresije

Drugi pristup u kreditnom skoringu jest pomoću linearne regresije. U ovom pristupu pokušava se pronaći najbolja linearna kombinacija karakteristika

$$w_0 + w_1X_1 + w_2X_2 + ... + w_pX_p = \mathbf{w}^* \mathbf{X}^{*\,T}$$

gdje je $\mathbf{w}^*$=($\mathrm{w}_0, w_1 ...., w_p$), $\mathbf{X}^*$=(1,$\mathrm{X}_1, X_2, ..., X_p$). Ako je p_i vjerojatnost da je podnositelj zahtjeva i u uzorku zadan, želimo pronaći $\mathbf{w}$ koji najbolje aproksimira

$$p_i = w_0 + x_{i1}w_1 + x_{i2}w_2 + ... + x_{ip}w_p, \quad \text{za sve } i. \tag{3.19}$$

Radi lakšeg snalaženja, pretpostavimo da je prvih n_G karakteristika u uzorku dobrih i njihova je vjerojatnost jednaka $p_i = 1$ za $i = 1....,n_G$, a idućih n_B karakteristika u uzorku je loših i njihova vjerojatnost iznosi $p_i = 0$ za $i = n_G + 1, ..., n_B$ i vrijedi $n_G + n_B = n$. U linearnoj regresiji tražimo koeficijent koji minimizira srednju kvadratnu pogrešku između lijeve i desne strane u (3.19). To je ekvivalentno minimiziranju izraza

$$\sum_{i=1}^{n_G}\left(1 - \sum_{j=0}^{p} w_j x_{ij}\right)^2 + \sum_{i=n_G+1}^{n_G+n_B}\left(\sum_{j=0}^{p} w_j x_{ij}\right)^2. \tag{3.20}$$

U vektorskoj notaciji, (3.19) može biti zapisan kao

$$\begin{pmatrix} 1 & \mathbf{X}_G \\ 1 & \mathbf{X}_B \end{pmatrix}\begin{pmatrix} w_0 \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix} \tag{3.21}$$

ili

$$\mathbf{Y}\mathbf{w}^T = \mathbf{b}^T \tag{3.22}$$

gdje je

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_G & \mathbf{X}_G \\ \mathbf{1}_B & \mathbf{X}_B \end{pmatrix}$$

$n \times (p + 1)$ matrica.

$$\mathbf{X}_G = \begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \\ x_{n_G 1} & \cdots & \cdots & x_{n_G p} \end{pmatrix}$$

je $n_G \times p$ matrica,

$$\mathbf{X}_B = \begin{pmatrix} x_{n_G+11} & \cdots & \cdots & x_{n_G+1p} \\ x_{n_G+21} & \cdots & \cdots & x_{n_G+2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_G+n_B 1} & \cdots & \cdots & x_{n_G+n_B p} \end{pmatrix}$$

je $n_B \times p$ matrica i vrijedi

$$\mathbf{b}^T = \begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix},$$

gdje je $\mathbf{1}_G(\mathbf{1}_B)$ $1 \times n_G(n_B)$ vektor sa svim jedinicama.

Da bismo pronašli koeficijente linearne regresije potrebno je minimizirati

$$(\mathbf{Y}\mathbf{w}^T - \mathbf{b}^T)^T(\mathbf{Y}\mathbf{w}^T - \mathbf{b}^T). \tag{3.23}$$

Izraz će biti minimiziran kad je derivacija jednaka nuli. To znači

$$\mathbf{Y}^T(\mathbf{Y}\mathbf{w}^T - \mathbf{b}^T) = 0 \text{ ili } \mathbf{Y}^T\mathbf{Y}\mathbf{w}^T = \mathbf{Y}^T\mathbf{b}^T,$$

$$\mathbf{Y}^T \cdot \mathbf{b}^T = \begin{pmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{X}_G & \mathbf{X}_B \end{pmatrix} \cdot \begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix} = \begin{pmatrix} n_G \\ n_G\mathbf{m}_G \end{pmatrix}$$

$$\text{i } \mathbf{Y}^T\mathbf{Y} = \begin{pmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{X}_G & \mathbf{X}_B \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{X}_G \\ \mathbf{1} & \mathbf{X}_B \end{pmatrix} = \begin{pmatrix} n & n_G\mathbf{m}_G + n_B\mathbf{m}_B \\ n_G\mathbf{m}_G^T + n_B\mathbf{m}_B^T & \mathbf{X}_G^T\mathbf{X}_G + \mathbf{X}_B^T\mathbf{X}_B \end{pmatrix}. \tag{3.24}$$

Ako smo radi objašnjenja koristili očekivanja uzorka kao stvarna očekivanja, dobivamo

$$\mathbf{X}_G^T\mathbf{X}_G + \mathbf{X}_B^T\mathbf{X}_B = n\mathbb{E}\{X_iX_j\} = n\,\mathrm{Cov}(X_i, X_j) + n_G\mathbf{m}_G\mathbf{m}_G^T + n_B\mathbf{m}_B\mathbf{m}_B^T$$

Ako je S kovarijacijska matrica uzorka, dobivamo

$$\mathbf{X}_G^T\mathbf{X}_G + \mathbf{X}_B^T\mathbf{X}_B = n\mathbf{S} + n_G\mathbf{m}_G\mathbf{m}_G^T + n_B\mathbf{m}_B\mathbf{m}_B^T \tag{3.25}$$

Koristeći (3.24) i (3.25) dobivamo

$$\begin{aligned} nw_0 + (n_G\mathbf{m}_G + n_B\mathbf{m}_B)\mathbf{w}^T &= n_G, \\ (n_G\mathbf{m}_G^T + n_B\mathbf{m}_B^T)w_0 + (n\mathbf{S} + n_G\mathbf{m}_G\mathbf{m}_G^T + n_B\mathbf{m}_B\mathbf{m}_B^T)\mathbf{w}^T &= n_G\mathbf{m}_G^T \end{aligned} \tag{3.26}$$

Zamjenjujući prvu jednadžbu u (3.26) drugom jednadžbom, dobivamo

$$\begin{aligned} ((n_G\mathbf{m}_G^T + n_B\mathbf{m}_B^T)(n_G - (n_G\mathbf{m}_G + n_B\mathbf{m}_B)\mathbf{w}^T)/n) \\ +(n_G\mathbf{m}_G\mathbf{m}_G^T + n_B\mathbf{m}_B\mathbf{m}_B^T)\mathbf{w}^T + n\mathbf{S}\mathbf{w}^T = n_G\mathbf{m}_G^T, \\ \text{tako je } \left(\frac{n_Gn_B}{n}\right)(\mathrm{m}_G - \mathbf{m}_B)\mathbf{w}^T + n\mathbf{S}\mathbf{w}^T = \left(\frac{n_Gn_B}{n}\right)(\mathbf{m}_G - \mathbf{m}_B)^T; \\ \text{tako je } \mathbf{S}\mathbf{w}^T = c(\mathbf{m}_G - \mathbf{m}_B)^T. \end{aligned} \tag{3.27}$$

(3.27) daje najbolji izbor $\mathbf{w} = (w_1, w_2, ..., w_p)$ kao koeficijente linearne regresije.

3.5 Logistička regresija

Regresijski pristup linearnoj diskriminaciji ima jedan očiti nedostatak. U (3.19) desna strana može poprimiti bilo koju vrijednost između $-\infty$ i $+\infty$, dok je lijeva strana vjerojatnost i poprima vrijednosti između 0 i 1. Bilo bi bolje kad bi lijeva strana bila funkcija od p_i što bi značilo da može poprimiti širi raspon vrijednosti. Tada ne bismo imali problem s tim da točke u podacima imaju vrlo slične vrijednosti kao zavisne varijable ili da regresijska

jednadžba predviđa vjerojatnosti manje od 0 ili veće od 1. Jedna takva funkcija mogla bi biti logaritam zadane vjerojatnosti p_i. To nas dovodi do logističkog regresijskog pristupa gdje je Wiginthon bio jedan od prvih koji ga je primijenio na rezultate kreditnog skoringa. U logističkoj regresiji, to bismo mogli zapisati kao

$$\log\left(\frac{p_i}{1-p_i}\right) = w_0 + w_1x_1 + w_2x_2 + ... + w_px_p = \mathbf{w}\cdot\mathbf{x}^T \tag{3.28}$$

Kako $\frac{p_i}{1-p_i}$ poprima vrijednosti između 0 i $+\infty$, funkcija $\log(\frac{p_i}{1-p_i})$ poprima vrijednosti između $+\infty$ i $-\infty$. Sada iz (3.28) slijedi

$$p_i = \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}} \tag{3.29}$$

To je pretpostavka logističke regresije. Zanimljivo je primijetiti, ako pretpostavimo da je distribucija dobrih i loših karakteristika višedimenzionalna normalna, tada taj primjer sadrži pretpostavku logističke regresije. Pretpostavimo još da je očekivanje za dobre karakteristike jednako μ_G i očekivanje za loše karakteristike jednako μ_B te da imaju zajednički kovarijacijsku matricu Σ. To znači da $\mathbb{E}(X_i|G) = \mu_{G,i}, \mathbb{E}(X_i|B) = \mu_{B,i}$ i $\mathbb{E}(X_iX_j|G) = \mathbb{E}(X_iX_j|G) = \Sigma_{ij}$. Odgovarajuća funkcija gustoće dana je s

$$f(\mathbf{x}|G) = (2\pi)^{-\frac{p}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp\left(\frac{-(\mathbf{x}-\mu_G)\Sigma^{-1}(\mathbf{x}-\mu_G)^T}{2}\right), \tag{3.30}$$

Ako je p_B proporcija populacije s lošim karakteristikama i p_G proporcija populacije s dobrim karakteristikama, tada je logaritam vjerojatnosti za klijenta i jednaka

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{p_Gf(\mathbf{x}|G)}{p_Bf(\mathbf{x}|B)}\right)$$
$$= \mathbf{x}\cdot\Sigma^{-1}2(\mu_B-\mu_G)^T + (\mu_G\cdot\Sigma^{-1}\cdot\mu_G^T + \mu_B\cdot\Sigma^{-1}\cdot\mu_B^T) + \log\left(\frac{p_G}{p_B}\right) \tag{3.31}$$

Budući da je to linearna kombinacija od x_i, ona zadovoljava pretpostavku logističke regresije.

U usporedbi s ostalim regresijama, poteškoća logističke regresije jest nemogućnost korištenja metode najmanjih kvadrata kako bi se izračunao koeficijent $\mathbf{w}$. U tom slučaju koristimo metodu maksimalne vjerodostojnosti (MLE ili maximum likelihood estimation).

Razlika između linearne regresije i logističke jest ta što linearna regresija pokušava uklopiti zadanu vjerojatnost p u danu linearnu kombinaciju atributa, dok logistička pokušava uklopiti $\log(\frac{p}{1-p})$.

Ostale dvije nelinearne funkcije koje se koriste u kreditnom skoringu jesu probit funkcija i stablo odlučivanja. O probit funkciji možete pročitati u [3, Poglavlje 4.6], a stabla odlučivanja objašnjena su u idućem poglavlju.

3.6 Stabla odlučivanja

Stabla odlučivanja jesu neparametarska tehnika klasifikacije klijenata u homogene skupine. Postoje dvije vrste čvorova u stablima odlučivanja:

krajnji čvor - njime završava određena grana stabla

čvor odluke - definira određeni kriterij u obliku vrijednosti atributa iz kojeg izlaze grane koje zadovoljavaju određene vrijednosti tog atributa.

Kod ove tehnike, skup odgovora dijeli se na dva podskupa. Kao što je već ranije objašnjeno, podnositelji zahtjeva označeni su kao dobri (nerizični) i loši (rizični) i cilj kreditnog skoringa jest pronaći klasifikatore koji najbolje razdvajaju dobre klijente od loših. Algoritam počinje podjelom glavnog skupa klijenata na dva poskupa koji sadrže uzorke dobrih, odnosno loših klijenata. Iz svake pojedine skupine podaci se dijele prema svim mogućim kriterijima ponovno u dvije grane. Pri tome se odabire kriterij koji podatke dijeli u skupine koje su više homogene od početne skupine podatka. Procedura se ponavlja sve dok podatke nije moguće dalje dijeliti u skupine koje su homogenije od početnih podataka. Tri su osnovna koraka u konstrukciji stabla odlučivanja:

1. izgradnja stabla odlučivanja s ulaznim podacima - čvorovima odluka, čvorovima posljedica, granama alternativnih akcija i granama posljedičnih stanja

2. računanje očekivanih vrijednosti odluka postupkom računanja unatrag - računanje započinje krajnjim čvorovima stabla i kreće se prema početnom čvoru odluke. Svakom čvoru pridružuje se očekivana vrijednost na način da je na završnom čvoru izračunata konačna vrijednost i alternativne, a čvoru posljedica pridružuju se očekivane vrijednosti izračunate prema formuli

$$\mathbb{E}V_{i-1} = \sum_j p_j \mathbb{E}V_i, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., m$$

gdje je $\mathbb{E}V_{i-1}$ očekivana vrijednost u čvoru $i-1$, p_j vjerojatnost grane j koja izlazi iz čvoja $i-1$. Na čvoru odluke, očekivana vrijednost jednaka je najvećoj vrijednosti izračunatoj algoritmom.

3. pronalaženje optimalnog puta postupkom računanja prema naprijed.

Za testiranje svakog pojedinog razdvajanja i mjerenja homogenosti podataka, koriste se Kolmogorov-Smirnovljeva statistika, indeks diverzifikacije, Gini indeks i indeks entropije.

Kolmogorov-Smirnovljeva statistika

Neka su $F(s|G)$ i $F(s|B)$ kumulativne funkcije distribucije dobrih tj. loših karakteristika X_i, gdje je s najbolja mjera za dobrog tj. lošeg klijenta. Ako je D dug nastao krivom kla-

sifikacijom lošeg klijenta kao dobrog, a L izgubljena zarada nastala krivom klasifikacijom dobrog klijenta kao lošeg, tada je potrebno minimizirati izraz

$$LF(s|G)p_G + D(1 - F(s|B))p_B \tag{3.32}$$

Ako je $Lp_G = Dp_B$, to je isto kao računati Kolmogorov-Smirnovljevu udaljenost između dvije distribucije tj. želimo minimizirati izraz $F(s|G) - F(s|B)$ ili maksimizirati izraz $F(s|B) - F(s|G)$. Ako dva podskupa podijelimo na lijevi (l) i desni (r) skup, onda je gornja maksimizacija jednaka maksimiziranju razlike $p(l|B) - p(l|G)$. Iz toga slijedi da je $p(l|B) = \frac{p(B|l)p(l)}{p(B)}$ prema Bayesovom pravilu. Tada Kolmogorov-Smirnovljev (KS) kriterij kaže da je potrebno pronaći raspodjelu na lijevi i desni skup i zatim maksimizirati

$$KS = |p(l|B) - p(l|G)| = \left| \frac{p(B|l)}{p(B)} - \frac{p(G|l)}{p(G)} \right| \cdot p(l). \tag{3.33}$$

Indeks diverzifikacije $i(v)$

Ako želimo podijeliti čvor v na lijevi l i desni r podčvor s vjerojatnostima $p(l)$ i $p(r)$ respektivno, može se odrediti mjera diverzije podjele tog čvora pomoću formule

$$I = i(v) - p(l)i(l) - p(r)i(r). \tag{3.34}$$

Najjednostavnije je, prilikom ove maksimizacije, definirati

$$\begin{aligned} i(v) &= p(G|v) \quad \text{za } p(G|v) \leq 0.5 \\ i(v) &= p(B|v) \quad \text{za } p(B|v) < 0.5. \end{aligned}$$

Gini indeks

Umjesto linearnog indeksa, Gini indeks je kvadratni i definiran je pomoću

$$i(v) = p(G|v)p(B|v),$$

i

$$G = p(G|v)p(B|v) - p(l)(G|l)p(B|l) - p(r)(G|r)p(B|r). \tag{3.35}$$

Indeks entropije

Još jedan nelinearni indeks jest indeks entropije, gdje je

$$i(v) = -p(G|v)\ln(p(G|v)) - p(B|v)\ln(p(B|v)). \tag{3.36}$$

Kao što samo ime kaže, to se odnosi na entropiju ili količinu informacija u podjeli čvora na dobar i loš podčvor.

Poglavlje 4

Bihevijoralistički kreditni skoring

4.1 Utjecaj PSD2 informacija na kreditni skoring i odluku o kreditiranju

U ovom ćemo poglavlju objasniti utjecaj direktive o izmijenjenim uslugama plaćanja (PSD2) na kreditni skoring i odluku o kreditiranju. Te su zaključke donijeli Domjan Barić, Marc Gaudart, Siniša Slijepčević i Toni Vlaić u svome znanstvenome radu na koji ćemo se pozvati tijekom ovog poglavlja.

Direktiva o izmijenjenim uslugama plaćanja poznatija kao PSD2 (The Revised Payment Services Directive) direktiva je Europske unije provedena početkom 2018. godine. Cilj te direktive jesu integriranje i optimiziranje europskog platnog tržišta, poticanje inovacija i konkurentnosti te unaprijeđivanje sigurnosti elektroničkog plaćanja. Omogućuje klijentima banke korištenje drugih poslužitelja usluga za obavljanje različitih aktivnosti na temelju svojih financijskih podataka. To zahtijeva od banaka davanje pristupa podacima o korisničkom računu ili pokreće plaćanje tim pružateljima usluga. Kao rezultat toga, korisnici bi trebali imati koristi od pristupa novim i inovativnim proizvodima i doživjeti poboljšane razine usluga.

Autori znanstvenoga rada pokazali su ovim projektom da se samo PSD2 podaci mogu koristiti za izvođenje izuzetno snažnog predviđanja propusta zaduživanja potrošača, koji se može koristiti za donošenje profitabilnih odluka o kreditiranju.

PSD2 u kreditiranju

PSD2 trebao bi promijeniti lanac vrijednosti plaćanja, profitabilnost nekih poslovnih modela maloprodajnog bankarstva i očekivanja kupaca. Postoje tri glavna pokretača direktive PSD2:

1. Poboljšanje prava potrošača i transparentnost

2. Poboljšanje sigurnosti putem SCA (snažna autentifikacija korisnika)
3. Omogućiti potrošačima jednostavno dijeljenje podataka o računu s trećim stranama

Kao posljedica toga, treće strane moći će, uz izričit dogovor s kupcima, pristupiti i koristiti podatke klijenata koji su dostupni na bankovnim računima. Ove informacije omogućit će veću razinu prilagodbe proizvoda.

Kreditiranje je značajan dio prihoda banaka i vrlo važnu ulogu u tome imat će uporaba podataka pomoću PSD2. Cilj projekta bio je istražiti na koji će način PSD2 utjecati na kreditiranje i što bi se moglo učiniti da se ta prilika iskoristi.

Istraživanje je pokazalo da podaci dobiveni pomoću PSD2 omogućuju zajmodavcima puno bolje razumijevanje razine rizika svakog klijenta. To omogućuje donošenje preciznijih odluka o kreditiranju, što rezultira personaliziranom ponudom proizvoda za kupca i dovodi do boljeg uzajamnog odnosa između banke i klijenta.

Poboljšano modeliranje rizika zahtijeva novi i drugačiji pristup modeliranju i prognoziranju kreditnog skoringa jer se temelji na različitim informacijama prikupljenim iz tradicionalnih modela. Varijable koje banke trenutno koriste u svojim modelima lako su razumljive.

Na primjer, često se koriste podaci o kreditnoj povijesti: Je li podnositelj zahtjeva za kredit kasnio s plaćanjem rate? Ako da, koliko puta? Koliko uzastopnih mjeseci?

Dostupne informacije razlikuju se od zemlje do zemlje, a kreditna povijest može uključivati različite korisne informacije - hipoteke, kredite za automobil, mobilni telefon i komunalne naknade. Zaključak je da se kreditnom poviješću, na ovaj način, određuje kreditna sposobnost. Podaci o plaćama i nekim vrstama transakcija na tekućem računu povremeno se uzimaju u obzir, ali ove varijable obično nisu glavni pokretači modela jer često zahtijevaju dodatne napore za izdvajanje.

Izazovi koje donosi korištenje PSD2

PSD2 pruža iscrpan prikaz prihoda i izdataka podnositelja zahtjeva tijekom vremena, u elektroničkom obliku. Postoje, međutim, tri glavna izazova koja trebaju biti riješena.

Dostupnost podataka o transakcijama na tekućem računu tijekom određenog vremenskog razdoblja stvara prvi izazov: stvaranje značajnih varijabli koje se mogu koristiti za donošenje kreditnih odluka. Skupovi podataka koji će biti dostupni putem sučelja PSD2 znatno se razlikuju od onih korištenih u tradicionalnim modelima. Tradicionalni modeli koriste značajke kao što je kreditna povijest ili prosječna plaća. Oni prikupljaju informacije iz više transakcija na strukturirani način. Međutim, glavni izazov pri razvoju modela kreditnog skoringa temeljen na PSD2 generiranim transakcijskim podacima jest da su dostupne samo značajke niske razine, kao što su pojedinačne transakcije. Tijekom projekta

model se temeljio na tisućama ulaznih podataka niske razine.

Drugi je izazov stvoriti univerzalni standard na području cijele Europske unije koji do sada ne postoji. Mnoge zemlje još uvijek su u procesu definiranja stvarnih tehničkih standarda koji će biti primjenjivi na određenom tržištu. To znači da bi modeli kreditnog skoringa koji su izgrađeni za bilo koje tržište trebali biti dovoljno fleksibilni da kako bi se mogli nositi s višestrukim budućim scenarijima s obzirom na dostupne informacije. Zadatak ovog projekta bio je definirati različite algoritame za svaki od zadanih scenarija.

Treći izazov bio je reducirati šumove i nepravilnosti koje nastaju tijekom vremena u transakcijama na tekućem računu. Te nepravilnosti otežavaju prepoznavanje određenih informacija kao što je plaća. Osim toga, podaci izdvojeni iz transakcija na tekućem računu u obliku specifičnih značajki ili varijabli izuzetno su korelirani. Ovo je značajan izazov za tradicionalne pristupe modeliranju kreditnog skoringa. Moderne tehnike strojnog učenja nude dobro rješenje za upravljanje tim izazovima.

Predvidljivost modela temeljenih na PSD2 sustavu

Autori su razvili i temeljito testirali modele strojnog učenja za bihevijoralistički kreditni skoring. Ti modeli predviđaju zadane postavke među maloprodajnim klijentima zasnovane samo na podacima dostupnim putem API-ja PSD2.

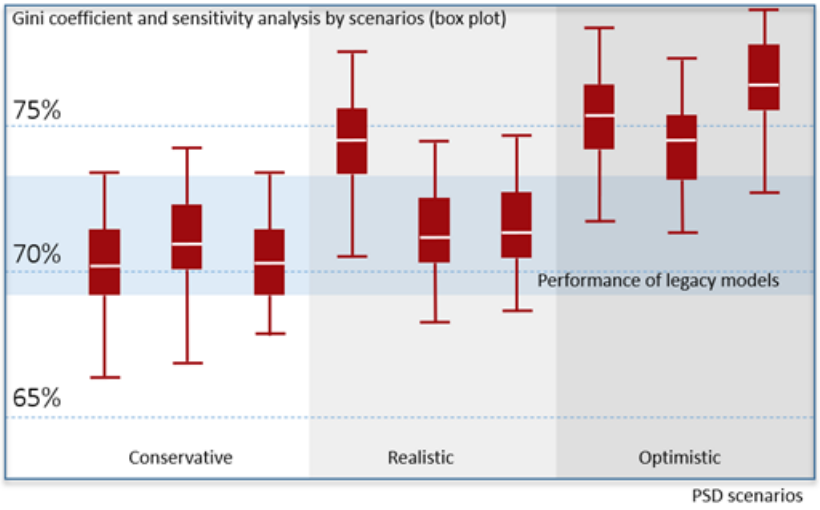
Metodologija razvoja modela

Da bi se razvili modeli, autori projekta radili su s jednom od vodećih banaka u Srednjoj Europi na uzorku od nekoliko milijuna klijenata. Modeli su razvijeni na svim aplikacijama za potrošačke kredite u razdoblju između svibnja 2016. i svibnja 2017. godine. Za izdvajanje financijskog ponašanja podnositelja zahtjeva korišteno je šest mjeseci transakcijske povijesti prije prijave za kredit.

Da bi se riješila nesigurnost s obzirom na to da će podaci biti dostupni putem PSD2, autori su razvili više modela - svaki korištenjem različitog skupa ulaznih podataka. Osnovni model koristi samo iznos i datum transakcije te salda tekućeg računa u trenutku transakcije. Gini indeks osnovnog modela u prosjeku iznosi 70%.

Najbolji (ali još uvijek realan) model ostvaruje Gini index, u prosjeku, od 76%. Ovaj model uključuje informacije o bilo kojem dostupnom prekoračenju na tekućem računu u trenutku transakcije i osnovne demografske podatke (adresu i dob).

Rezultati za 9 modeliranih scenarija prikazani su na Slici 4.1.



Slika 4.1: Prikaz razvijenih modela u 9 različitih scenarija, mjerenih pomoću Gini indeksa, Izvor: [5]

Izbor tehnika u strojnom učenju

Autori projekta upotrijebili su nekoliko algoritama strojnog učenja kako bi pronašli najbolji. Isprva su razvijali jednostavne modele kao što su linearna i logistička regresija. Ovi jednostavniji modeli imali su Gini indeks u rasponu od 55% do 59%. Iz ove relativno čvrste izvedbe jednostavnijih modela zaključak je da su nastale varijable zabilježile osnovne osobine kupaca. Nadalje, testirani su složeniji algoritmi poput umjetnih neuronskih mreža. No ti su algoritmi puno osjetljiviji na korelaciju i zato se ne koriste u daljnjem radu.

Izrada značajki

Ključna prednost ovakvog pristupa bila je metodologija kojom su autori izdvojili podatke iz transakcijskih podataka. Stvorili su oko 3.000 kompozitnih značajki u različitim vremenskim razdobljima, koja obuhvaćaju različite trendove i događaje. Koristili su različite matematičke i statističke tehnike za generiranje tih značajki.

Značajke uključuju sljedeće primjere:

(i) Razne statističke značajke u određenim vremenskim razdobljima, uključujući minimum, maksimum, očekivanje, standardnu devijaciju i slično

(ii) Fourierove transformacije

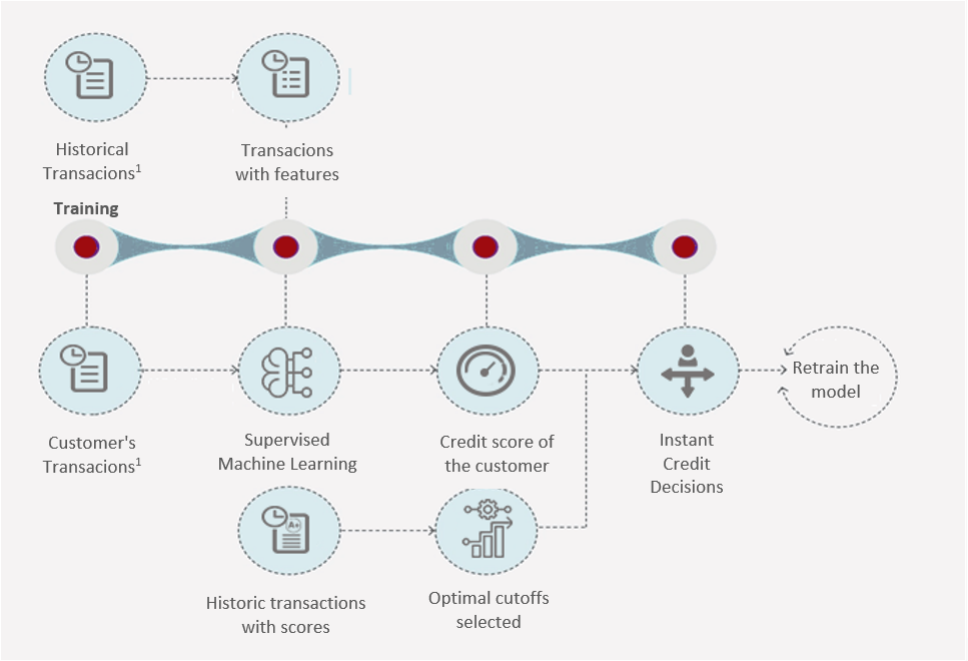
(iii) Značajke izračunate po prilagođenim algoritmima za procjenu plaće, stabilnost plaće, broj propuštenih kreditnih isplata itd.

(iv) Varijable učestalosti određenih događaja, npr. plaćanja u određenom rasponu.

Proces projekta

Razvijen je jednostavan algoritam koji je izračunao plaću korisnika pomoću analize kreditnih transakcija tijekom vremena. Nakon što su se autori uvjerili da je njihov kalkulaor za plaće točan, razvili su novu značajku pod nazivom stabilnost plaće. Značajka stabilnost plaće tijekom projekta pokazala se kao važnija značajka u odnosu na samu plaću.

Iznos i vrsta potrošnje bile su važne značajka, no trendovi i varijabilnost potrošnje tijekom vremena pružili su dodatne informacije koje su bile snažni prediktori kreditnog rizika. Takve kompozitne osobine tada su bile unesene u nadgledani model strojnog učenja kao varijable kako bi se povećala prediktivnost. Analizirano je nekoliko različitih scenarija, razvijeno više različitih modela na slučajno odabranim podatcima i testirani su algoritmi strojnog učenja kako bi se osiguralo postizanje najboljih mogućih rezultata. Naposljetku su autori surađivali s klijentom kako bi implementirali razvijeni model kreditnog skoringa unutar vlastitih IT sustava i procesa. To je prikazano na slici ispod.



Slika 4.2: Informacije dostupne putem sučelja kompatibilnog s PSD2, Izvor: [5]

Zaključak rada

S PSD2 promjenama, banke će imati pristup punom profilu klijenta (uključujući račune u drugim bankama). Kao rezultat toga, oni će moći prilagoditi svoje proizvode potrebama svojih kupaca kroz suvremenu tehnologiju. Kako bi se iskoristila ova prilika, potrebno je koristiti tehnologije poput strojnog učenja jer se time ostvaruje konkurentnost na tržištu koje se brzo mijenja. Autori rada bili su u mogućnosti stvoriti predvidljive modele za kreditni skoring koristeći samo one podatke dostupne putem PSD2. Pokazali su da je moguće razumjeti cjelokupno financijsko ponašanje pomoću transakcijskih podataka. Pokazali su i da je moderni pristup strojnom učenju znatno prediktivniji od tradicionalnog pristupa (temeljenog na regresijskim tehnikama). Ovim pristupom smanjuje se vrijeme čekanja na odobrenje kredita i troškovi obrade zahtjeva postaju manji.

Bibliografija

[1] Dr. Nikola Sarapa. *Teorija vjerojatnosti*. Školska knjiga, Zagreb, 1987.

[2] Robert W. Keener *Theoretical Statistics - Topics for a Core Course.* Springer Science+Business Media, New York, 2009.

[3] Lyn C. Thomas, David B. Edelman, Jonathan N. Crook *Credit Scoring and Its Applications.* Society for Industrial and Applied Mathematics, Philadelphia, 2002.

[4] Dr. Ljiljanka Kvesić *Primjena stabla odlučivanja u kreditnom skoringu.* Ekonomski vjesnik, God XXVI.,BR. 2/2013. str. 382-390

[5] Domjan Barić, Marc Gaudart, Siniša Slijepčević, Toni Vlaić *PSD2 information has the power to transform credit scoring and lending decisions* U pripremi.

Sažetak

Kreditni skoring numerički je sustav pomoću kojeg se ocjenjuje rizičnost klijenta kojemu se želi prodati neki proizvod. U ovome radu proučavamo statističke metode za izračun kreditnog skoringa. Na kraju opisujemo rezultate primjena modernih statističkih metoda i mašinskog učenja na uspješnu izradu vrlo prediktivnog behavijoralnog kreditnog skoringa.

Summary

Credit scoring is a numerical system used to assess the risk of a customer to whom a product is sold. In this thesis we consider statistical methods for calculating the credit scoring. Finally, we describe results of an application of modern statistical methods and machine learning to a successful development of very predictive behavioral credit scoring.

Životopis

Rođena sam 31. kolovoza 1994. godine u Našicama, malom gradiću u srcu Slavonije. Odrasla sam u Koški, gdje sam pohađala osnovnu školu Ivane Brlić-Mažuranić i gdje je moja draga profesorica Tomica Vujić u meni potaknula ljubav prema matematici. Uz njezinu podršku upisujem III. gimnaziju u Osijeku. Profesor Ilija Ilišević priprema me za natjecanja i daljnje školovanje kroz izborni predmet matematike.

Tijekom osnovnoškolskog i srednjoškolskog obrazovanja sudjelujem na natjecanjima iz logike, informatike i matematike gdje postižem rezultate čak i na državnim natjecanjima.

2013. godine upisujem Preddiplomski sveučilišni studij matematike na Matematičkom odsjeku Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. 2016. godine završavam Studij i stječem titulu univ.bacc.math. Iste godine upisujem Diplomski studij Matematička statistika.

Tijekom studija zapošljavam se u Erste banci kao mlađi programer u data warehouse timu. Uz to, tijekom cijelog studija volonterski držim besplatne instrukcije iz matematike za osnovnu i srednju školu u crkvi Sveta Mati Slobode.